2019

Detecting Ancient Balancing Selection: Methods And Application To Human

Katherine Siewert
University of Pennsylvania, kmsiewert@gmail.com

Detecting Ancient Balancing Selection: Methods And Application To Human

Abstract
Balancing selection can maintain genetic variation in a population over long evolutionary time periods. Identifying genomic loci under this type of selection not only elucidates selective pressures and adaptations but can also help interpret common genetic variation contributing to disease. Summary statistics which capture signatures in the site frequency spectrum are frequently used to scan the genome to detect loci showing evidence of balancing selection. However, these approaches have limited power because they rely on imprecise signatures such as a general excess of heterozygosity or number of genetic variants. A second class of statistics, based on likelihoods, have higher power but are often computationally prohibitive. In addition, a majority of methods in both classes require a high-quality sequenced outgroup, which is unavailable for many species of interest. Therefore, there is a need for a well-powered and widely-applicable statistical approach to detect balancing selection. Theory suggests that long-term balancing selection will result in a genealogy with very long internal branches. In this thesis, I show that this leads to a precise signature: an excess of genetic variants at near identical allele frequencies to one another. We have developed novel summary statistics to detect this signature of balancing selection, termed the β statistics. Using simulations, we show that these statistics are not only computationally light but also have high power even if an outgroup is unavailable. We have derived the variance of these statistics, allowing proper comparison of β values across sample sizes, mutation rates, and allele frequencies - variables not fully accounted for by many previous methods. We scanned the 1000 Genomes Project data with β to find balanced loci in humans. Here, I report multiple balanced haplotypes that are strongly linked to both association signals for complex traits and regulatory variants, indicating balancing selection may be affecting complex trait architecture. Due to their high power and wide applicability, the β statistics enable evolutionary biologists to detect targets of balancing selection in a range of species and with a degree of specificity previously unattainable.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Benjamin F. Voight

Keywords
Balancing selection, Natural selection, Selection scans

Subject Categories
Bioinformatics | Evolution | Genetics

DETECTING ANCIENT BALANCING SELECTION: METHODS AND APPLICATION TO HUMAN

Katherine M. Siewert

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania
in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
2019

Supervisor of Dissertation

Benjamin F. Voight, Ph.D., Associate Professor of Systems Pharmacology and Translational Therapeutics and Genetics

Graduate Group Chairperson

Benjamin F. Voight, Ph.D., Associate Professor of Systems Pharmacology and Translational Therapeutics and Genetics

Dissertation Committee:
Marcella Devoto, Ph.D., Professor of Pediatrics and Epidemiology
Sarah A. Tishkoff, Ph.D., Professor of Genetics and Biology
Junhyong Kim, Ph.D., Professor of Biology
Philipp W. Messer, Ph.D., Assistant Professor of Biological Statistics and Computational Biology, Cornell University

Acknowledgments

I am lucky enough to have a large number of people to thank for making my time in graduate school both enjoyable and fruitful. I would first and foremost like to thank my advisor, Ben Voight. He is in large part responsible for turning me into the scientist I am today, both by modeling science as a process driven by excitement and curiosity and through viewing it as a question-driven endeavor. His support, trust and patience allowed my PhD to be a happy six years. In addition, I am grateful for my thesis committee: Marcella Devoto, Sarah Tishkoff, Junhyong Kim and Philipp Messer for their time and gracious advice.

I have been lucky to have been a part of the amazing group of individuals that compose the Voight Lab. Kelsey Johnson, you deserve an especially large shout-out, as it has been both fun and motivating working towards our PhDs together. Thank you for all of our conversations, whether it be about science or life, as well as for the occasional well-deserved sass when I said something truly ridiculous! Thank you to Paul Babb for being a constant source of support and encouragement, and for being a model of hard and careful work throughout my PhD. I must also thank Varun Aggarwala for

encouraging my inquisitive and outspoken nature, even to the occasional exasperation of our poor advisor and labmates.

Kat Gawronski – I thank you for being both a close friend and an amazing support system – conversations with you seemed to either lead to laughter, me learning something interesting, or both! Onur Yörük, the same goes for you. I would also like to thank Kim Lorenz, Chris Thom, and Diana Cousminer for their wise feedback and support throughout my PhD. Finally, I would like to thank Will Bone and Chris Adams for making my last few years of PhD full of fun discussion, both scientific and otherwise.

I owe much to the inspiring and supportive scientific community at Penn, including my friends, professors, administrators and colleagues. I would like to give a special acknowledgment to the GCB students in my year: Alex Amlie-Wolf, Lucy Shan, Salika Dunatunga, Brett Beaulieu-Jones and Ian Mellis. I could not have asked to be a part of a more intelligent, mature and supportive group of individuals.

I also thank my family for turning me into the person I am today. I am grateful for my parents for modeling and encouraging curiosity, critical thinking and hard work, for which I credit my PhD to more than anything else. Gretchen Siewert, you are a never-ending source of laughter and support! Finally, I would like to thank my partner Jason Rocks. You've been my rock throughout graduate school and consistently inspire me to be the best scientist I can be.

DETECTING ANCIENT BALANCING SELECTION: METHODS AND APPLICATION TO HUMAN

Katherine M. Siewert

Benjamin F. Voight, Ph.D.

Balancing selection can maintain genetic variation in a population over long evolutionary time periods. Identifying genomic loci under this type of selection not only elucidates selective pressures and adaptations but can also help interpret common genetic variation contributing to disease. Summary statistics which capture signatures in the site frequency spectrum are frequently used to scan the genome to detect loci showing evidence of balancing selection. However, these approaches have limited power because they rely on imprecise signatures such as a general excess of heterozygosity or number of genetic variants. A second class of statistics, based on likelihoods, have higher power but are often computationally prohibitive. In addition, a majority of methods in both classes require a high-quality sequenced outgroup, which is unavailable for many species of interest. Therefore, there is a need for a well-powered and widely-applicable statistical approach to detect balancing selection. Theory suggests that long-term balancing selection will result in a genealogy with very long internal branches. In this thesis, I show that this leads to a precise signature: an excess of genetic variants at near identical allele frequencies to one another. We have developed novel summary statistics to detect this signature of balancing selection, termed the β statistics. Using simulations, we show that these statistics are not only

computationally light but also have high power even if an outgroup is unavailable. We have derived the variance of these statistics, allowing proper comparison of β values across sample sizes, mutation rates, and allele frequencies - variables not fully accounted for by many previous methods. We scanned the 1000 Genomes Project data with β to find balanced loci in humans. Here, I report multiple balanced haplotypes that are strongly linked to both association signals for complex traits and regulatory variants, indicating balancing selection may be affecting complex trait architecture. Due to their high power and wide applicability, the β statistics enable evolutionary biologists to detect targets of balancing selection in a range of species and with a degree of specificity previously unattainable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Overdominance is "due to the occurrence of a rather special class of mutations and gene combinations, which confer on heterozygotes a higher adaptive value... Although overdominance is, by and large, an exceptional situation, it is of particular interest to a student of population genetics".

— Theodosius Dobzhansky, 1952

1.1 A brief history of balancing selection

1.1.1 Development of the theory of overdominance as a source of hybrid vigor

The concept of balancing selection arose from early discussions of hybrid vigor. This phenomenon had been observed for centuries and has been of significant interest due to its direct relevance to plant breeding (Crow, 1987). Indeed, it was noted by Mendel, who observed that hybrid pea strains were larger and more vigorous than parental strains (Mendel 1865). Charles Darwin also had an interest in the topic: he wrote an entire book on inbreeding depression and hybrid vigor (Darwin, 1878). Geneticists in the early 1900s, particularly George Shull and Edward East, suggested that this hybrid vigor, or heterosis, was due to the increased diversity of alleles found in an individual with higher heterozygosity. It was proposed that these alleles increase fitness in a complementary fashion to one another (East, 1936; Shull, 1948). The term overdominance was introduced by Fred Hull to refer to this phenomena. He defined it as the situation in which the fitness of a heterozygote would be over the fitness that would be observed if either allele was dominant (Hull, 1945, 1946). Although overdominance fell out of favor as the reason for hybrid vigor (see section 1.4.1), it continued to be of general interest as a possible selective force acting on genomes.

1.1.2 Modern definitions of balancing selection

Throughout the next several decades, balancing selection became defined as natural selection in which multiple alleles are maintained at a locus in a population (Levene, 1953). It can be due to overdominance as originally suggested, but further work demonstrated that it can also be due to spatially, temporally, or negative-frequency dependent selection. For instance, if multiple niches are present in an environment an equilibrium can occur if alternate alleles are beneficial in the different niches (Levene, 1953; Haldane, J.B.S., Jayakar, 1963). Furthermore, if the fitness of an allele in a population fluctuates with time, then under certain conditions, this can lead to long-term maintenance of the alternately favored alleles (Hedrick *et al.*, 1976). Finally, the fitness of an allele may be inversely proportional to its frequency, which will cause the frequency of the allele to increase until it is no longer favored and has therefore reached its equilibrium frequency (Takahata and Nei, 1990).

1.1.3 Examples of balanced loci

Throughout the last century, there has been an interest in finding genomic loci that have experienced balancing selection. There are several classic sites long proposed to be under this type of selection. Perhaps the most famous is the Hemoglobin-β locus. Homozygotes for the sickle-cell allele have sickle-cell anemia, homozygotes for the other alleles have an increased risk of malaria, while heterozygotes have resistance to malaria and at most have a mild case of sickle-cell (Luzzatto, 2012; Aidoo *et al.*,

2002). The major histocompatibility complex (MHC) region has also been long hypothesized to be under selection for multiple reasons (Slade and McCallum, 1992). The first is overdominance, as it could be advantageous for an immune system to be able to respond to a wider diversity of pathogens. However, studies have shown that the level of heterozygosity observed in the MHC in humans cannot be explained solely by overdominance (De Boer *et al.*, 2004). Frequency-dependent selection may be responsible for the additional signal of balancing selection in the MHC. The mechanism for this type of selection would be that pathogens may not be adapted to overcome rare human alleles that aid in resistance against them (Slade and McCallum, 1992).

There have also been a number of loci recently proposed to be under balancing selection with experimental or observational evidence (Schweizer *et al.*, 2018; Sano *et al.*, 2018; Network, 2015; Wheat *et al.*, 2010). One example is balancing selection on a locus in North American wolves. Homozygotes and heterozygotes for the K_B allele have a black coat color, while homozygotes for the k^y allele have a gray coat color (Anderson *et al.*, 2009). Interestingly, heterozygotes have the highest fitness in Yellowstone populations, suggesting that coat color is not the only selective pressure (Schweizer *et al.*, 2018). Evidence suggests that overdominance may be acting at this locus, possibly due to the K locus being involved with not only coat color, but also immune response (Schweizer *et al.*, 2018).

Another recently described example is spatially-dependent selection in a species of extremophile cyanobacterium. A polymorphism which affects the function of heterocysts, which are nitrogen-fixing cells, has been maintained for tens of millions of years

in this species (*Fischerella thermalis*) and has significantly different allele frequencies between individuals living in two different temperatures. There is high gene flow between the individuals living in the different temperatures, and very low population differentiation elsewhere in the genome (Sano *et al.*, 2018). This suggests that these individuals are part of the same species and that this locus may be under long-term spatially dependent selection due to adaptation to different temperature conditions.

1.2 The effect of balancing selection on coalescence and patterns of variation

1.2.1 Effects of balancing selection on the coalescent process

Initially, a newly balanced allele will increase in frequency. This creates long haplotypes of limited diversity, mimicking the effects of an incomplete positively selected sweep (Charlesworth, 2006). The allele will then increase in frequency until it reaches what is termed its equilibrium frequency – the frequency at which it is expected to be maintained. This frequency is determined by the relative fitness of the different genotypes. For instance, if the fitness of the two homozygote classes are equal, and the fitness of the heterozygote is higher, then the equilibrium frequency will be 50%. In the case of the sickle cell allele in populations in malaria-endemic regions, the homozygotes for the sickle-cell alleles have much lower fitness than for the opposite allele. This low fitness results in the allele frequency of the sickle-cell allele being

much lower than 50%. In malaria endemic regions estimates have found its frequency to be no more than 18% in any population (Piel *et al.*, 2010).

If the selective pressure is sustained, then balancing selection can maintain alleles in populations for potentially very long time periods, given certain conditions. Specifically, the selective coefficient must be high enough that the heterozygotes have a significant fitness advantage over homozygotes (Robertson, 1962). In addition, the equilibrium frequency must be of intermediate frequency (between about 20 and 80%) (Takahata and Nei, 1990; Ewens and Thomson, 1970; Robertson, 1962), or genetic drift will remove the variation after enough generations.

By maintaining polymorphism, balancing selection affects the structure of the coalescent tree at the locus. Under neutrality, genetic drift will cause coalescence of all lineages after a moderate amount of time. In contrast, neither allele can fix in the population under balancing selection, so the time to most recent common ancestor (TMRCA) will predate the start of balancing selection, making it potentially much older than at a neutral locus **(Fig. 1.1)** (Kaplan *et al.*, 1988). The genealogy of each allelic class, defined as all haplotypes containing one of the two balanced alleles, will be nearly identical to that of a neutral locus of sample size equal to the size of the allelic class (Hey, 1991). The number of individuals in the two sub-trees is determined by the equilibrium frequency. Although these characteristics will hold true under all types of long-term balancing selection, the structure of the coalescent tree under temporally-dependent selection may be more complex, due to the relative sizes of the allelic classes varying throughout time.

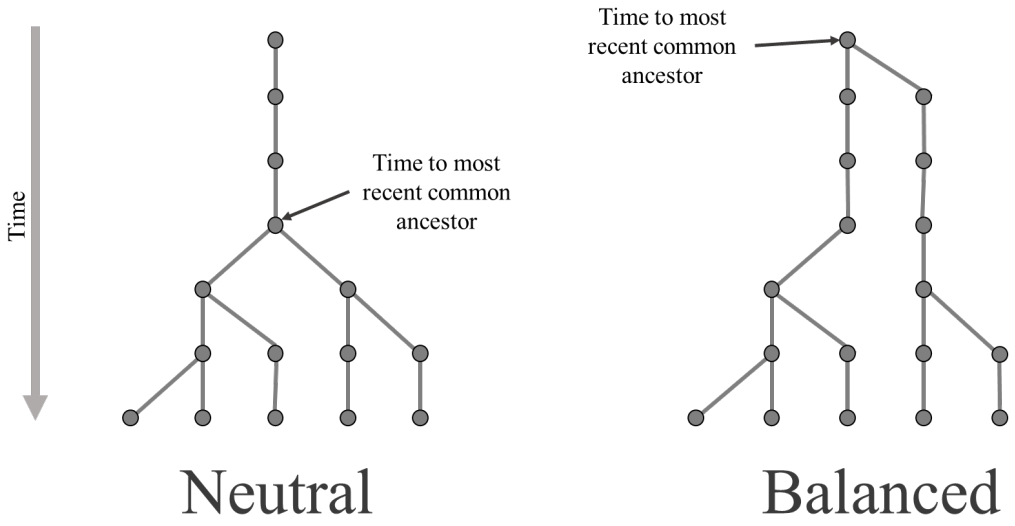

Figure 1.1: Balancing selection increases the time to most recent common ancestor at a locus. Circles represent haploid individuals.

1.2.2 Effects of extended time to most recent common ancestor on the site frequency spectrum

The long time to most recent common ancestor at a locus under balancing selection results in old haplotypes. Due to their age, these haplotypes have had time to accumulate large numbers of mutations (Charlesworth 2006). More specifically, balanced haplotypes will accumulate their own unique alleles, but these alleles are not allowed to fix in the population because selection constrains the frequency of the haplotype class in which they arose (Hey, 1991). This results in the classic signature of balancing selection: an excess number of intermediate frequency alleles and a deficit of substitutions (i.e. genomic positions in which the allele in all ingroup individuals differs from the outgroup individual) **(Fig. 1.2)** (Hudson *et al.*, 1987; Tajima, 1989).

The effect of balancing selection on the coalescent process is analogous to a hap-

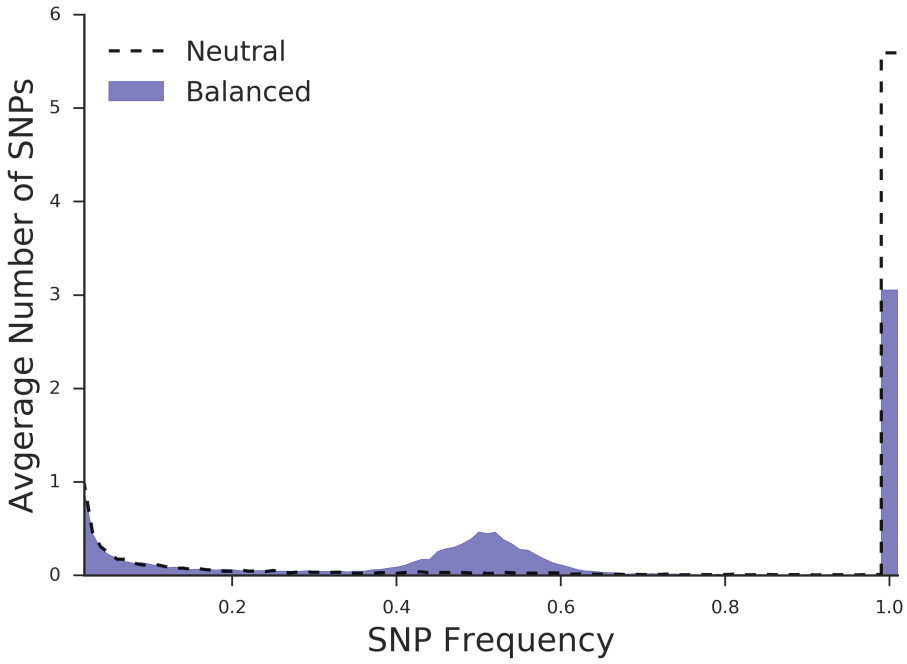

Figure 1.2: Site frequency spectrum of derived alleles in balanced or neutral simulations, with core variant removed. Substitutions, *i.e.* positions in which the derived allele is fixed in the species under consideration, are displayed as SNPs of frequency 1.0. Window size is 500 base pairs on either side of the core site, with sample size 100 chromosomes. Based on simulations with an equilibrium frequency of 0.5.

loid two-island model (Hey, 1991). In this model, a population is split into two isolated subpopulations. Mutations can arise on each island, but without migration between the islands, the mutations will not reach the subpopulation on the other island. Instead, these mutations build-up on the islands, causing an excess number of intermediate frequency alleles when the sub-populations are combined into a site frequency spectrum (Tajima, 1989). However, migration will allow alleles to transfer between the two islands, reducing the number of unique alleles. Analogously, under balancing selection, two haplotype classes are maintained in the population with neither one allowed to fix due to selection. Mutations unique to each allelic class will build-up throughout time. Eventually, recombination will occur, decoupling the

mutations from the effects of selection (Hudson and Kaplan, 1988; Hey, 1991).

1.3 Detecting balancing selection: statistics and scans

1.3.1 Motivation for detecting balancing selection

A number of fundamental questions in evolutionary biology can be addressed through scans for natural selection. One key question is what selective pressures species have experienced throughout their evolutionary history, and how they have adapted to these pressures. If a balanced locus is detected in a scan for selection, then computational and/or experimental approaches may be used to figure out what phenotypes the locus is associated with. In some cases, the causal selective pressures on the locus can be inferred. This process has successfully uncovered multiple targets of balancing selection and their associated phenotypes, as discussed in section 1.1.3, though I note that only some of these began with a genome-wide scan for selection. Scans for positive selection and follow-up have been more successful, possibly owing to a larger history of methodological development for detecting positive selection. Established sites under positive selection in humans with an established phenotype include EDAR for hair follicle thickness (Kamberov *et al.*, 2013), lactase persistence (Bersaglieri *et al.*, 2004; Tishkoff *et al.*, 2007), alcohol dehydrogenase (Osier *et al.*, 1999; Whitfield, 2002), and selection on *PDE10A* for spleen size in sea nomads (Ilardo

et al., 2018). These successes in scans for positive selection bode well for the goal of detecting and explaining sties under balancing selection.

In fact, one advantage of detecting balancing selection is that it leaves a much narrower footprint in the genome than does positive selection (**Section 2.4**). This results in a smaller number of possible causal variants, making the identification of the true causal variant easier. Despite this factor, the number of balanced loci in humans with an established phenotype and/or selective pressure is very limited, motivating the need research on balancing selection.

A second key question scans for selection can help answer is the impact different types of selection have had on patterns of variation in species. This is discussed more in section 1.4. In short, in order to identify the prevalence of balancing selection, we must first develop a better understanding of its effects on genomic loci under this type of selection.

In order to answer both these questions, a highly specific signature of balancing selection, and a corresponding high-powered test for its detection, is needed.

1.3.2 Classic methods for detecting balancing selection based on the site frequency spectrum

By scanning population-level sequencing data for the signatures of selection, loci which have experienced long-term balancing selection can be detected. Current meth-

ods of doing so calculate a statistic sensitive to the effects of balancing selection on the site frequency spectrum in a sliding window across the genome.

Tajima's D is one such statistic. It is the difference of two unbiased estimators of the mutation rate. Intuitively, these estimators estimate the mutation rate by counting the number of SNPs, using the intuition that a higher mutation rate will result in a higher number of mutations in a window. Accordingly, estimators of the mutation rate will be higher if there are more SNPs. The first estimator which comprises Tajima's D, θ_π, estimates the mutation rate using heterozygosity. Because the number of intermediate frequency mutations on old haplotypes is expected to be higher than on newer haplotypes **(Fig. 1.2)**, this estimator will increase in windows which have experienced long-term balancing selection. The second estimator, θ_W, uses the total number of SNPs in a window to estimate the mutation rate. This estimator is relatively insensitive to balancing selection and is used to correct for the background mutation rate. Tajima's D is the difference of these two estimators divided by the standard deviation (Tajima, 1989):

$$D = \frac{\theta_\pi - \theta_W}{Var[\theta_\pi - \theta_W]} \tag{1.3.1}$$

Therefore, values of D significantly above zero indicate potential long-term balancing selection, while values close to zero indicate an absence of evidence of balancing selection.

Another commonly used statistic, the Hudson-Kreitman-Aguad (HKA) test, does not

look at allele frequencies, but instead compares only the number of polymorphisms and the number of substitutions to their expected number under neutrality. By combining these terms in a chi-squared statistic, significant deviation from neutrality can be detected (Hudson *et al.*, 1987). Specifically, a higher number of polymorphisms, and a lower number of substitutions are expected under balancing selection, as previously discussed.

Several other statistical tests have also been used to detect these signatures of balancing selection. The Mann-Whitney U test can be used to detect an excess number of intermediate-frequency alleles. This test can be used in combination with a modified HKA test, which detects an excess number of variants at a locus. The union of these tests produced a set of loci with both higher-frequency SNPs and a higher total number of SNPs than the background levels in the human genome, indicating balancing selection (Andres *et al.*, 2009).

1.3.3 Trans-species SNPs and haplotypes as a signature of balancing selection

An orthogonal signature of balancing selection is shared SNPs or haplotypes between multiple species. Trans-species haplotypes are defined as two or more variants that are found in tight linkage disequilibrium and are shared between humans and a primate outgroup (in our case, chimpanzee). If a neutral SNP was present in a common ancestor to two species, under most conditions it is expected to have drifted out of

the population in one or both species, leading to a substitution. In contrast, if the SNP is under balancing selection in both species, it can be maintained from the time it arose until present (Takahata, 1990; Takahata and Nei, 1990). This leads to the segregation of both alleles in both species. Therefore, if two species share one SNP (a trans-species SNP) or more than one SNP at a locus (a trans-species haplotype), this indicates potential balancing selection.

The presence of trans-species SNPs may be due to recurrent mutations (i.e. the same mutation occurs in both species independently), so are not a test for selection with high specificity. In contrast, due to the very low probability of two recurrent mutations occurring in high linkage disequilibrium under human and chimp demography (Gao *et al.*, 2014), trans-species haplotypes are a very specific signature of balancing selection in humans. Multiple studies have used human and chimp sequencing data to detect these shared SNPs and haplotypes (Leffler *et al.*, 2013; Teixeira *et al.*, 2015). These scans have identified a number of balanced loci potentially involved in immunity, including loci involved with recognizing *plasmodium falciparum* (Leffler *et al.*, 2013) or a missense change in LAD1, an autoantigen which causes linear IgA disease (Teixeira *et al.*, 2015). In addition, the ABO blood group has been proposed to be under long-term balancing selection in humans on the basis of trans-species comparisons (Ségurel *et al.*, 2012; Teixeira *et al.*, 2015).

1.3.4 Recent statistics to detect balancing selection: Composite likelihood methods

An ideal statistic to detect balancing selection would be a full likelihood estimation of a locus being under balancing selection as opposed to being neutral. Such a statistic could be based on summary level information about the site frequency spectrum, such as the probability of seeing a mutation at each frequency at each distance from a balanced SNP. Alternatively, it could make use of individual-level genotype data, calculating the likelihood of the observed haplotype structures at various distances from the balanced SNP. These are in contrast to the early statistics designed to detect balancing selection, which do not rely on likelihoods and instead use a summary statistic to capture the general patterns caused by selection.

Recently, two composite likelihood methods were developed to detect balancing selection (DeGiorgio *et al.*, 2014) which utilize the site frequency spectrum. These statistics are composite in that they consider each SNP independently of the other SNPs at the locus. The Kaplan-Darden-Hudson model, which describes the genealogy of a neutral SNP linked to a selected SNP (Kaplan *et al.*, 1988; Hudson and Kaplan, 1988), is used to model the probability of seeing a segregating site or substitution at each recombination-scaled distance from a balanced SNP. The background levels of polymorphism and substitutions are used to generate the expected site frequency spectrum near a SNP evolving neutrally. By comparing these two composite likelihoods, a test for balancing selection, T1, is derived (DeGiorgio *et al.*, 2014). The $T1$ statistic looks only at the presence of polymorphisms and substitutions but does not

consider allele frequencies.

DeGiorgio et al. then derive the $T2$ statistic, which does take into account allele frequencies (DeGiorgio *et al.*, 2014). However, because the site frequency spectrum under balancing selection is unknown, simulations are used to generate probabilities of seeing SNPs at various frequencies. These simulations are performed under specified parameter values, including a large range of equilibrium frequencies and recombination rates. By using simulations matched for equilibrium frequency and recombination rate at a locus, these simulation-generated likelihoods are incorporated into a composite likelihood framework.

1.3.5 Power and applicability of existing method for detecting balancing selection

Using simulations, DeGiorgio *et al.* (2014) demonstrated that the power of their $T1$ and $T2$ statistics are higher than the HKA test and Tajimas D. The power of their $T2$ method is higher than $T1$, as would be expected because it considers allele frequencies. However, $T2$ presents challenges in its applicability. Namely $T2$, like the $T1$ and HKA test, requires an outgroup with which to call substitutions and ancestral/derived allele states. In addition, prior to scanning the genome with the $T2$ test, large numbers of simulations must be performed to generate expected site frequency spectra. These simulations are computationally intensive, making wide applicability of the $T2$ statistic difficult.

In some cases, a sequenced individual from an outgroup species is unavailable, rendering all these summary statistics inapplicable except Tajima's D. However, Tajima's D has the lowest power to detect balancing selection in the analysis of DeGiorgio *et al.* (2014). Furthermore, calling trans-species SNPs and trans-species haplotypes requires multiple outgroup individuals. This suggests the need for new methods to detect selection which do not require an outgroup but have the high power of the $T2$ method.

1.3.6 Coalescent methods

Recent methods seek to directly estimate the time to most recent common ancestor (TMRCA), as opposed to using summary statistics. These methods are based on the pairwise sequentially Markovian coalescent (PSMC) method (Li and Durbin, 2011). This method models coalescent times between two individuals at a locus along the genome using a hidden Markov model, with the hidden state being the TMRCA, and emissions being whether the two haploid individuals match (produce a homozygote) or have different alleles (produce a heterozygote). Transitions between states are the result of recombination. The longer the coalescence time between the individual, the more time there has been for mutations to build-up between them. Therefore, the number of heterozygote sites will be proportional to the TMRCA of the individuals at the locus. Specifically, the number of SNPs occurring on a branch is exponentially distributed with rate equal to the individual mutation rate multiplied by the branch

lengths.

More recent methods have adapted the PSMC method to multiple genomes. One such method, ARGweaver, was used to detect loci with extremely old TMRCAs, indicative of balancing selection (Rasmussen *et al.*, 2014). However, these methods are computationally intensive, taking multiple days to weeks to estimate genome-wide TMRCAs on a high-powered computer. Recent methods have continued to improve on these methods, both in terms of speed and applicability, however, they remain prohibitively computationally expensive for general use (Palamara *et al.*, 2018; Speidel *et al.*, 2019).

1.4 Genome-wide impact of balancing selection

1.4.1 Debate on the importance of balancing selection to evolution

Since its conception, there has been an unanswered question about prevalence of balancing selection in the evolutionary history of both humans and other species. In the early to mid 1900s, there was a debate as to why the increased number of heterozygotes seen in hybrid individuals increases vigor. Many argued it was due to there being less recessive deleterious alleles in the homozygote state in hybrids, termed the dominance hypothesis (Bruce, 1910). Others favored the idea that it was due to overdominance, supported by the view at the time that there was a higher number of

mutations in populations than would be expected, as would be expected due to long term balancing selection (Crow, 1998). However, the overdominance explanation fell out of favor, as it became clear that the mutation rate was higher than previously thought and that some of the observed overdominance was due to deleterious recessive alleles being linked to vigor-increasing dominant alleles (Moll *et al.*, 1963). In addition, experimental evidence showed that the dominance hypothesis better fit the fitness patterns seen with various genetic crosses (Crow, 1998).

However, despite the general consensus that overdominance was not as widespread as previously thought, the field still lacked an understanding of exactly how rare it was. The availability of genome sequencing from humans and other primates allowed a reconsideration of the debate decades later. An early scan for trans-SNPs using expressed sequence tags and virtual transcripts found little evidence of trans-species SNPs between human and chimpanzee (Asthana *et al.*, 2005). A year later, a scan for high polymorphism density found no loci showing significant evidence of ancient balancing selection (Bubb *et al.*, 2006).

However, more recent datasets, which contain whole-genome, high-quality genetic variation data, enable a more comprehensive look into the prevalence of ancient balancing selection. Multiple recent papers have looked for shared haplotypes between human and one or more primate outgroups and have found a number of shared haplotypes. Leffler *et al.* (2013) found 125 shared haplotypes between human and chimpanzee. A more recent paper looked for trans-species SNPs shared between humans, chimpanzees and bonobos and found 4 trans-species SNPs after performing

very conservative filters (Teixeira *et al.*, 2015)

These trans-species SNPs and haplotypes are potentially only a small number of the total balanced loci in the genome. This is because for a balanced locus to be trans-species, the balancing selection must predate speciation, and the balanced haplotypes must not have drifted out of the population in either species, which could occur either because of a change in selective pressures or demography encouraging loss of variation. Therefore, the presence of trans-species SNPs and haplotypes in the genome indicate that balancing selection may have played a larger role in the evolution of humans than previously thought. However, the extent to which balancing selection has shaped patterns of variation in humans remains an open debate (Hedrick, 2012; Key *et al.*, 2014).

1.4.2 Effects of balancing selection on the deleterious mutation load

One might predict that deleterious mutation which occur in a species will be quickly removed by purifying selection. Therefore the number of deleterious mutations should be low, and any deleterious mutations that do segregate should be of recent origin and at low frequency. In contrast to this expectation, it has been suggested that there is an excess number of intermediate frequency deleterious mutations segregating in the human genome, termed the deleterious mutation load (Henn *et al.*, 2015). One reason for this might be human demographic parameters which make purifying selection less

effective (Keinan and Clark, 2012; Eyre-Walker and Keightley, 1999). However, there is ongoing debate as to how much of the deleterious mutation load can be credited to human demography (Do *et al.*, 2015; Simons *et al.*, 2014).

An alternative explanation for the deleterious mutation load in humans is balancing selection, which can increase the deleterious mutation load via multiple mechanisms. The first is that the deleterious mutation can be the direct target of balancing selection, as is the case with the sickle cell allele at the human hemoglobin-β locus (Allison, 1954). The second is that the deleterious mutation can be on the same haplotype as the sweeping allele upon the start of balancing selection. The deleterious mutation will be swept up to intermediate frequency along with the balanced haplotype and will be maintained in the population until being decoupled from the balancing selection due to recombination. It has been proposed that this mechanism is responsible for some fraction of the deleterious mutations in the MHC region (Lenz *et al.*, 2016). Therefore, if balancing selection is common throughout the genome, it could be partly responsible for the deleterious mutation load in humans.

By increasing the number and frequency of deleterious variants, balancing selection may raise the heritability of complex traits by increasing the variance in the trait explained by genetics. This leads to the untested hypothesis that balanced loci may have increased trait heritability. Furthermore, if this hypothesis is true, then balanced loci should be prioritized in scans for disease-causing loci, as they have a higher probability of causing disease a priori.

1.5 Motivation for a new method for detecting ancient balancing selection

In summary, understanding where balancing selection has acted on the genome is of interest for multiple reasons: (1) it can reveal selective pressures, (2) adaptations to those pressures, (3) identify loci which may be influencing risk for disease and (4) help explain the deleterious mutation load. However, high power and widely-applicable methods for detecting balancing selection are critical to answer all four of these questions. As discussed in section 1.3, prior to this thesis, methods to detect this type of selection suffered from at least one of the following drawbacks: (1) They were of lower power, (2) they required an outgroup sequence or (3) they were too computationally intensive for wide applicability. It is the aim of this thesis to develop a method without these shortcomings.

Chapter 2

Detecting ancient balancing selection using an excess of allele frequency similarity

The results of this chapter are presented in:

Siewert, K.M. and Voight B.F. 2017. Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Molecular Biology and Evolution*, 34(11): 2996 3005.

2.1 Effects of balancing selection on the site frequency spectrum

2.1.1 A forward in time perspective

Consider a new neutral mutation that arises within an outcrossing, diploid population. In a genomic region not experiencing selection, this mutation is expected to eventually either drift out of the population, or become fixed (i.e., become a substitution). However, if the SNP is under balancing selection, then the allele's frequency can reach no higher than the frequency of the balanced allele it arose in linkage with, assuming no recombination. This is because the frequency is constrained by selection. Without a recombination event and given enough time, variants that are fixed within these allelic classes (defined by the selected variant) accumulate **(Fig. 2.1)**. In addition to this build-up of variants, there will be a corresponding reduction in the number of substitutions, because the variants that may have fixed in the population without balancing selection can instead reach a frequency no higher than that of the balanced allele that they are linked to.

2.1.2 A coalescent perspective

As discussed prior (Section 1.2), balancing selection causes long internal branches on a locus's coalescent tree. These internal branches will be ancestral to all sampled individuals in an allelic class, but not ancestral to individuals in the other allelic

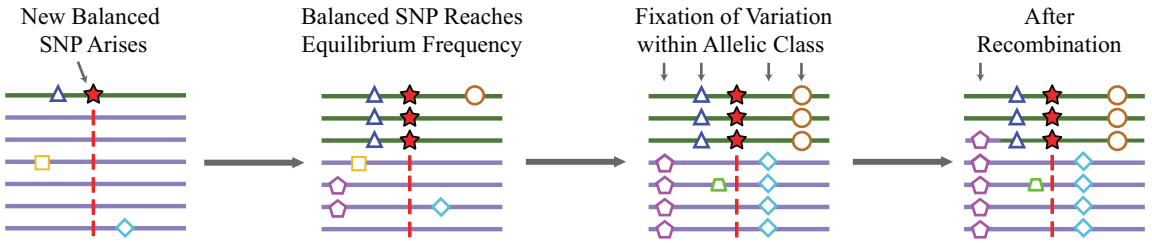

Figure 2.1: Model of allelic class build-up. (1) A new SNP (red star) arises in the population and is subject to balancing selection. (2) It sweeps up to its equilibrium frequency. (3) New SNPs enter the population linked to one of the two balanced alleles and some drift up in frequency. However, unlike in the neutral case, their maximum frequency is that of the balanced allele they are linked to, so variants build-up at this frequency (e.g., blue diamond or brown circle). (4) Recombination decouples SNPs (e.g., purple pentagon) from the balanced site, allowing them to experience further genetic drift.

class. Therefore, mutations occurring on these branches will be fixed within their allelic class (i.e. at the frequency of the balanced allele that they are linked to) **(Fig. 2.2)**. This contrasts with a tree representing a neutral locus, in which all lineages will have coalesced more recently. Any mutations occurring on the tree after (going backwards in time) this coalescent event will have occurred in an ancestor to all individuals at this locus and will therefore be a substitution when the locus is compared to an outgroup species. Therefore, once again, our model of balancing selection predicts that under balancing selection there will be an excess number of variants at identical frequency to the balanced alleles and a deficit of substitutions, when compared to the neutral model.

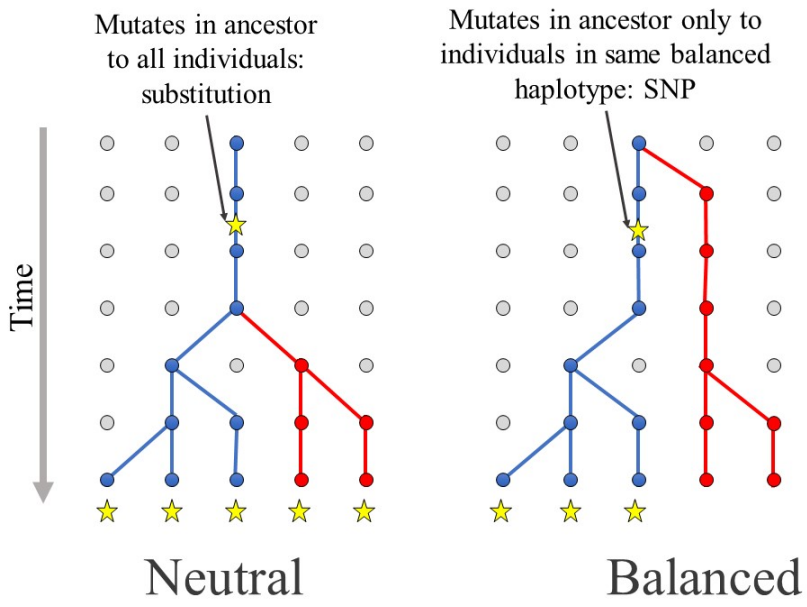

Figure 2.2: Long internal branches cause build-up of alleles at identical frequencies under balancing selection. Branches are colored by allelic class, which here have frequencies 3/5 (blue) and 2/5 (red).

2.1.3 Effects of recombination on the signature of balancing selection

Eventually, recombination decouples variants from the balanced allele, which allows them to drift to loss or fixation within the population **(Fig. 2.1)**. However, even after recombination, the frequency of the genetic variants previously fixed in their allelic class will remain close to that of their previous class until enough time has passed for genetic drift to significantly change their frequencies. In our simulations of balancing selection, a window expected to have experienced recombination since selection's start still has an excess number of variants at similar frequencies to the balanced variant. However, there is a smaller excess at identical frequencies relative to the narrower window, demonstrating the effects of recombination **(Fig. 2.3)**.

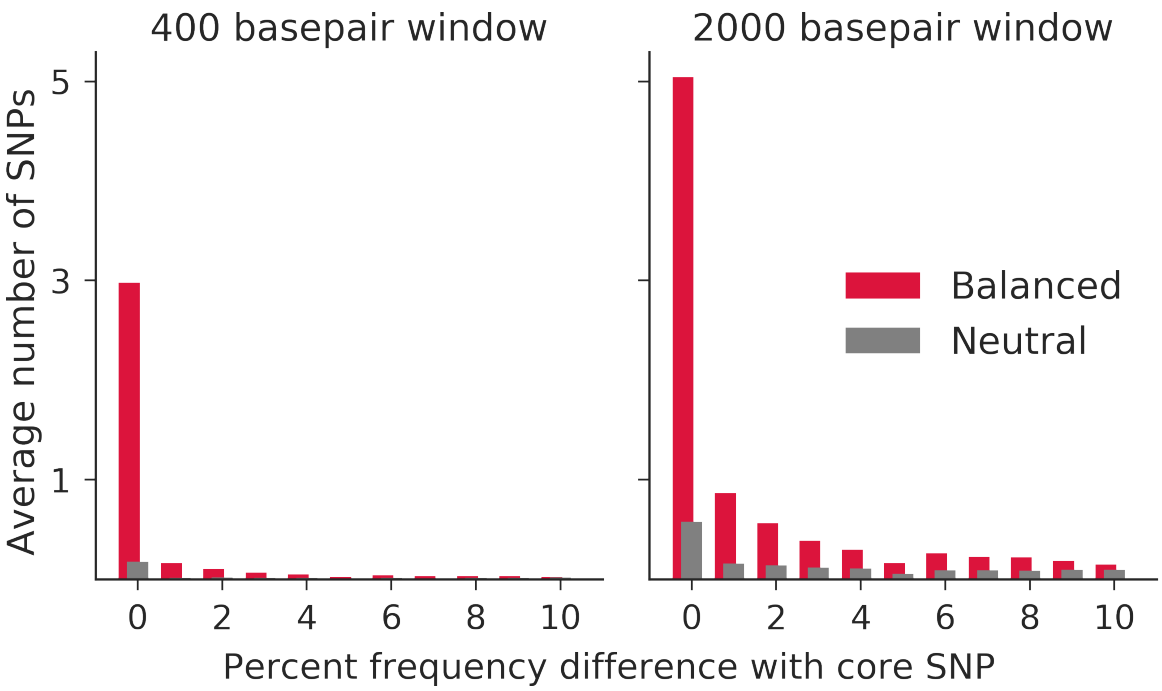

Figure 2.3: Simulations demonstrate the build-up of alleles at frequencies similar to balanced alleles as compared with selectively neutral counterparts. The 400 basepair window is not expected to have experienced recombination between allelic classes since the start of selection, whereas the 2000 basepair window is more likely to have.

2.2 The $\beta^{(1)}$ statistics for detecting balancing selection

2.2.1 Framework for capturing excess allele frequency correlation

To detect loci under ancient balancing selection we therefore want to develop a summary statistic which captures an excess number of SNPs at near identical frequencies to one another. We will use several components to do this. The first is a measure of allele frequency similarity. This allows one to weight SNP counts based on their frequency similarity to a core SNP. Next, we incorporate this measure of similarity

into an estimator of a mutation rate. Finally, we use this estimator in combination with another estimator that measures the background mutation rate as our statistic. In addition, we derive the variance of this statistic, so we can standardize it.

In this chapter I present two versions of this statistic. The first, the unfolded version, takes into account the ancestral/derived state at each SNP. By doing so, it can give more weight to SNPs of higher frequency, because they are less likely under neutrality. The second version, the folded version, does not use allele ancestral/derived states. Therefore, it is applicable even to species without a high-quality sequenced outgroup species.

2.2.2 Capturing allele frequency correlation

To capture allele frequency correlation, we derive a measurement of frequency similarity between a core variant and a second variant of interest. Let n be the number of chromosomes sampled, f_0 be frequency of the core SNP, f_i be the frequency of the second SNP, i, and p be the scaling constant. Finally, $g(f)$ returns the folded allele frequency and m is the maximum possible folded allele frequency difference between the core SNP and SNP i, We then measure the similarity in frequency, d_i, by:

$$g(f) = \min(f, n - f) \tag{2.2.1}$$

$$m = \max\Big(g(f_0), \frac{n}{2} - g(f_0)\Big) \tag{2.2.2}$$

$$d_i = \left(1 - \frac{|g(f_0) - g(f_i)|}{m}\right)^p \tag{2.2.3}$$

Thus, $g(f_0) - g(f_i)$ is the folded frequency difference between the core SNP and the SNP under consideration. We then divide by m, the maximum folded frequency difference possible with the core SNP, to get the percent of the maximum frequency different the two SNPs have. We then take 1 minus the result to give a similarity metric instead of a distance metric. We raise it to the power p so that we can weight variants in a non-linear fashion with respect to this fraction. Therefore, d_i can range from 0 if a SNP has the maximum frequency difference with the core SNP, to 1 if SNP i is at the same frequency as the core SNP **(Fig. 2.4)**. Guidance on the choice of p is given in section 2.2.3. We use the folded site frequency spectrum in calculating d_i, as the frequency difference between the core variant and the second variant is independent of whether the derived or ancestral allele of the nearby allele is in linkage with the derived or ancestral core allele.

In a region under long-term balancing selection, the average d_i between a core SNP and the surrounding variants is expected to be elevated. However, d_i alone is not optimally powered to detect balancing selection, as its value will be sensitive to changes in the mutation rate in the surrounding region, and it does not take into account the probability of seeing each allele frequency under neutrality.

2.2.3 Choice of p parameter

The power of our method lies in capturing allele frequency correlations. The parameter p controls how similar of allele frequencies to the core site are captured. As p

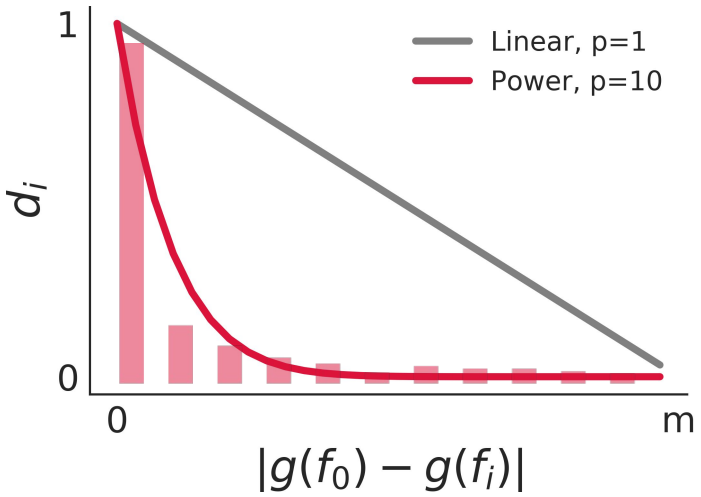

Figure 2.4: Absolute value of allele frequency similarity with core SNP ($|g(f_0)-g(f_i)|$) versus allele frequency similarity (d_i) as used by the β statistics, by different values of the p parameter. The grey and red lines represent the value of d_i at the given frequency similarity, while the light red bars represent the number of SNPs at a given frequency difference away from the core SNP in simulations of balancing selection, based on the 2000 basepair window panel of Fig. 2.3.

approaches infinity, the only sites that contribute towards θ_B are those that exactly match the frequency of the core SNP. At $p = 0$, all SNPs contribute the same amount to the estimate of $\hat{\theta}_B$, and so $\hat{\theta}_B$ becomes equivalent to $\hat{\theta}_w$. Simulations show that our method is fairly robust to choice of p **(Fig. 2.5)**.

That said, the optimal p will depend on the data set at hand. If allele frequency estimates are known to be inaccurate or sample sizes vary between SNPs, then a lower p may be more optimal, because variants fixed in allelic class may not accurately be called as being at identical frequency to the core SNP. In addition, by including SNPs at very similar frequency to the core SNP in the calculation of $\hat{\theta}_B$, SNPs that were once fixed in class, but are no longer due to recombination followed by a small amount of drift, are included. However, making p too low will result in the inclusion of allele frequencies that are very different than the balanced allele's frequency **(Fig. 2.4)**.

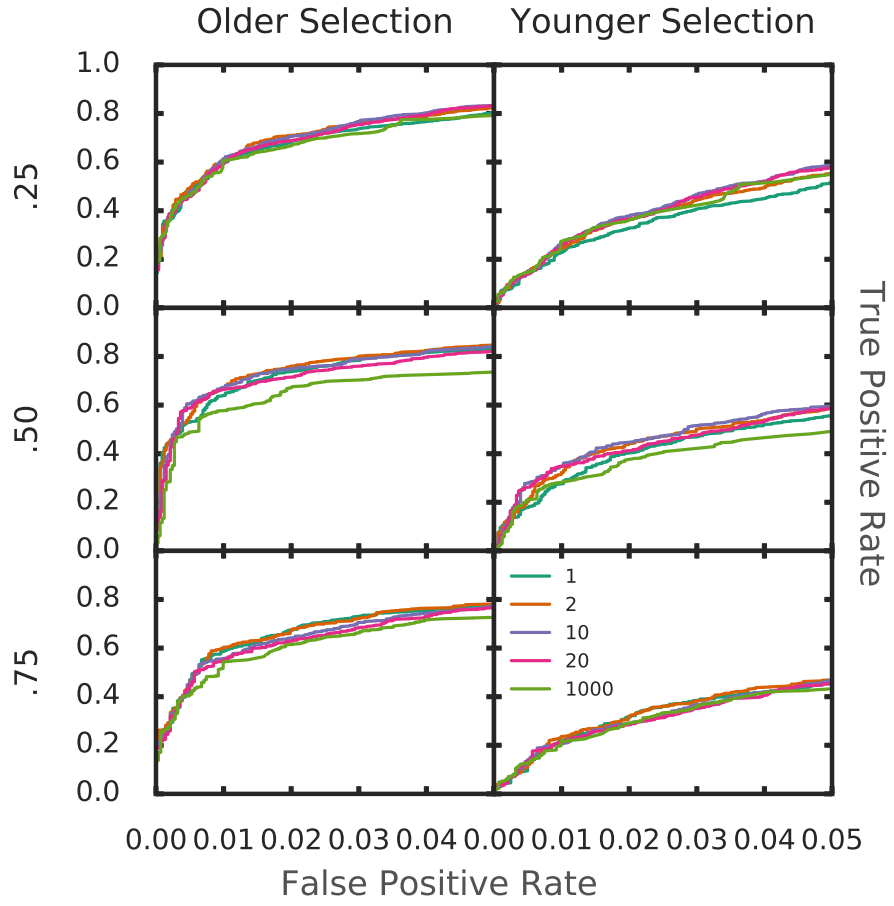

Figure 2.5: Power of methods to detect ancient balancing selection using different value of p parameter with Beta

In this chapter, we chose a $p = 20$, which gives the most weight to exact frequency matches, and a small amount of weight to very near, but not exact frequencies. If varying sample sizes are used for each SNP, then a lower p value may be optimal **(Fig. 2.21)**.

2.2.4 Estimator of the mutation rate based on allele frequency correlation

Derivation of Unfolded θ_B

Let n be the number of chromosomes sampled, d_i be the similarity measure and S_i be the number of variants at frequency i in the sample.

$$E[\sum_{i=1}^{n-1} id_i S_i] = \sum_{i=1}^{n-1} E[id_i S_i] \tag{2.2.4}$$

$$= \sum_{i=1}^{n-1} id_i E[S_i] \tag{2.2.5}$$

$$= \sum_{i=1}^{n-1} id_i \frac{1}{i}\theta \tag{2.2.6}$$

$$\hat{\theta}_\beta = \frac{\sum_{i=1}^{n-1} id_i S_i}{\sum_{i=1}^{n-1} d_i} \tag{2.2.7}$$

Derivation of Folded θ_B

$$E[\sum_{i=1}^{n-1} d_i S_i] = \sum_{i=1}^{n-1} E[d_i S_i] \tag{2.2.8}$$

$$= \sum_{i=1}^{n-1} d_i E[S_i] \tag{2.2.9}$$

$$= \sum_{i=1}^{n-1} d_i \frac{1}{i} \theta \tag{2.2.10}$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n-1} d_i S_i}{\sum_{i=1}^{n-1} d_i \frac{1}{i}} \tag{2.2.11}$$

Let $g(x)$ be the folded frequency of a SNP of frequency x, $S_{g(x)}$ be the number of SNPs at that folded frequency, $h = .5(n-1)$ and $m = .5n$. Folding the site frequency spectrum, we obtain:

$$\hat{\theta}^*_\beta = \frac{\sum_{i=1}^{m} d_i S_{g(i)}}{\sum_{i=1}^{m} d_i (\frac{1}{i} + \frac{1}{n-i}) \frac{1}{1+\delta_{i,n-i}}} \tag{2.2.12}$$

2.2.5 A summary statistic to detect balancing selection based on the site frequency spectrum

We propose a statistic, β, that uses our measure of allele frequency correlation, d_i, incorporated in θ_β, combined with a measure of the overall mutation rate, to detect balancing selection. Our approach is inspired by previous summary statistics of the site frequency spectrum (Tajima, 1989; Fay and Wu, 2000). These methods compute

the difference between two estimators of θ, the population mutation rate parameter, one of which is more sensitive to characteristics of the site frequency spectrum distorted in the presence of natural selection. We propose to calculate β at each SNP in a region of interest to identify loci in which there is an excess of variants near the core SNP's allele frequency, as evidence of balancing selection.

It has been previously shown that the mutation rate in a region can be estimated as: $\hat{\theta}_i = S_i * i$, where S_i is the total number of derived variants found i times in the window from a sample of n chromosomes in the population (Fu, 1995). An estimator of θ can then be obtained by taking a weighted average of θ_i. In our method, we weight by the similarity in allele frequency to the core SNP, as measured by d_i. If there is an excess of variants at frequencies close to the core SNP allele frequency, then our new estimator, θ_β, will be elevated. We propose:

$$\beta^{(1)} = \hat{\theta}_\beta - \hat{\theta}_w \tag{2.2.13}$$

$$\beta^{(1)*} = \hat{\theta}^*_\beta - \hat{\theta}_w \tag{2.2.14}$$

θ_w is simply Watterson's estimator (Watterson, 1975). β is, in effect, a weighted sum of SNP counts based on their frequency similarity to the core SNP. We exclude the core site from our estimation of θ_w and θ_β.

Under neutrality, the expected value of β is zero, because it is a difference of two unbiased estimators. In contrast, under balancing selection it is expected that there

will be an excess number of SNPs at near identical frequencies to one another elevating θ_β substantially over the true mutation rate in the window, while θ_W will be elevated only slightly. Therefore, values of β significantly above zero are suggestive of long-term balancing selection.

2.2.6 Properties of $\beta^{(1)}$ in simulations

To better understand the properties of β, we used simulations (for details see section 2.5.1) to examine its distribution with and without a balanced SNP.

As expected, under long-term balancing selection β tends to be greater than 0, and under neutrality it tends to be close, but slightly higher than, 0 **(Fig. 2.6)**. Under neutrality it is not exactly zero, because all the neutral windows β is actually calculated on will have at least one SNP, and the site frequency spectrum conditioned on seeing a SNP of a certain frequency does not equal the unconditional site frequency spectrum, as discussed in section 2.2.7 (Ferretti *et al.*, 2018).

We note that the mean value of β in our neutral simulations generally increases slightly with higher equilibrium frequencies. This behavior is expected because higher frequency alleles will tend to have a longer TMRCA and therefore higher diversity. The exception to this trend is neutral SNPs of frequency 0.5, which we posit is due to the fact that this allele frequency requires the most time for mutations to drift up to the equilibrium frequency needed to fix in their allelic class.

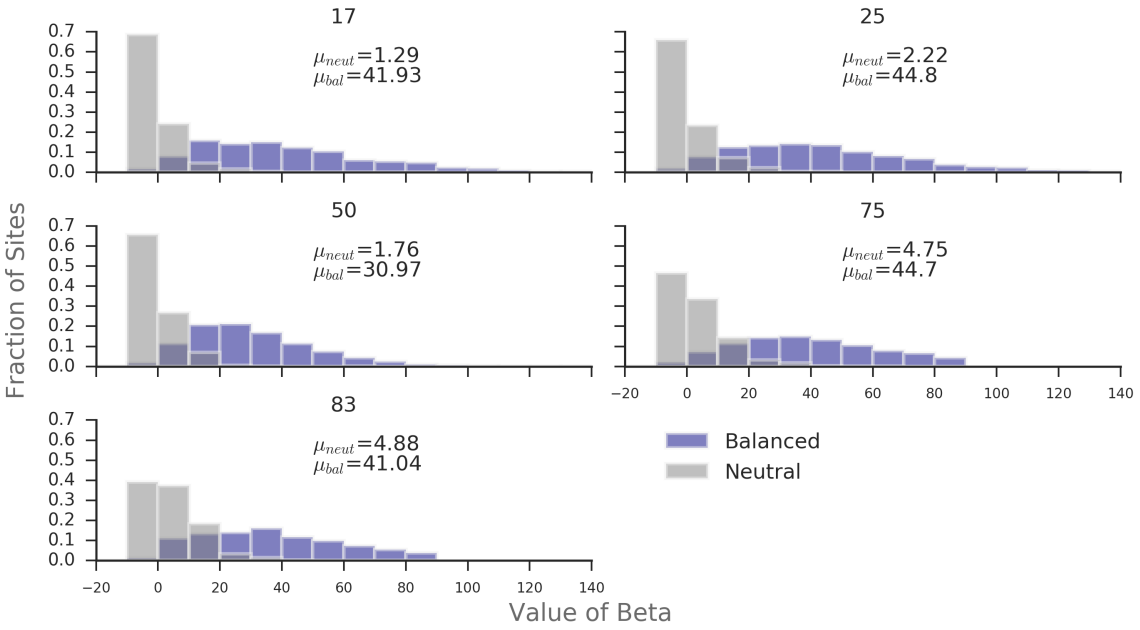

Figure 2.6: Distribution of $\beta^{(1)}$ in 1kb windows around a core SNP at different equilibrium frequencies. Based on simulations using default parameters. μ refers to the mean value of $\beta^{(1)}$ in balanced or neutral simulations.

2.2.7 On the assumption of independence between basepairs

In our derivations of $\hat{\theta}_\beta$ we do not use the conditional site frequency spectrum. In other words, the formula we use for the expected value and variance in SNP counts does not take into account the frequency of the core site. However, the conditional and unconditional SFS are unequal, as conditioning on the core SNP's frequency gives some knowledge about which underlying tree structures are most likely. Recently, two papers deriving the moments of the conditional SFS were published (Ferretti *et al.*, 2018; Klassmann and Ferretti, 2018). We used these moments to derive a modified $\hat{\theta}_\beta$ and $\hat{\theta}_W$ conditioned on the core SNP being at the observed frequency. However, doing so decreased power **(Fig. 2.7)**. We posit that this is because under the conditional site frequency spectrum, the expected number of SNPs at identical frequency to the

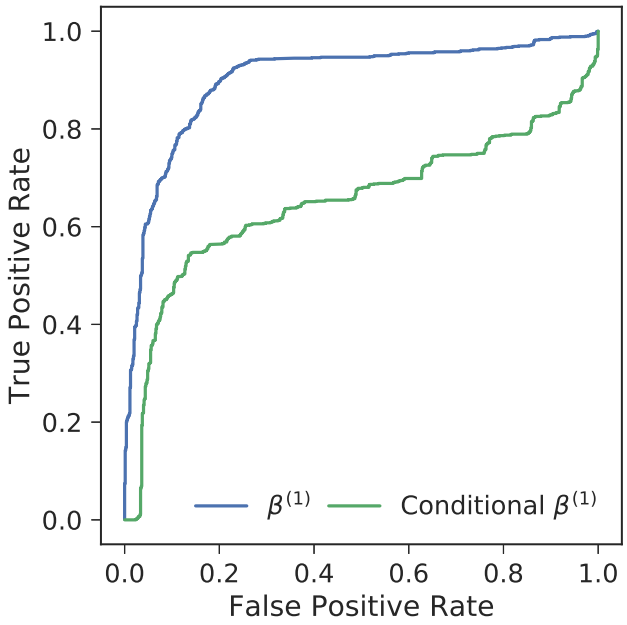

Figure 2.7: Power of $\beta^{(1)}$ statistic when derived using the unconditional site frequency spectrum of Fu (1995) versus conditional of Ferretti *et al.* (2018).

core SNP is increased. Therefore, when an estimator of the mutation rate is derived using this expected value, each SNP at that frequency is weighted less than if using an unconditional site frequency spectrum. This behavior is opposite the ideal: we want to give the most weight to SNPs at identical frequency to the core SNP, not less. Therefore, the power of this statistic is reduced, so we use moments of the unconditional site frequency spectrum to derive our β statistics and their variances.

2.3 Standardization of the $\beta^{(1)}$ statistics

We next derive the variance of our statistics, enabling normalization of β. This allows β scores to be properly compared across a range of parameters which can affect its

distribution, including population size, survey sample size, equilibrium frequencies, and mutation rate. This is a feature lacking from other summary statistics, with the exception of Tajima's D (Tajima, 1989).

2.3.1 Variance of the unfolded β statistic

The $Var[\hat{\theta}_\beta]$ can be obtained from the formula for variance of a general group of estimators presented in (Achaz, 2009) for which $\hat{\theta}_\beta$ is a member. σ is defined in Achaz (2009) and d_i is the measure of frequency similarity presented in section 2.2.2.

$$Var[\hat{\theta}_\beta] = \Big(\sum_{i=1}^{n-1} d_i\Big)^{-2}\left(\theta\Big(\sum_{i=1}^{n-1} d_i^2 i\Big) + \theta^2\Big(\sum_{i=1}^{n-1} d_i^2 i^2 \sigma_{ii} + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n-1} ij d_i d_j \sigma_{ij}\Big)\right) \tag{2.3.1}$$

2.3.2 Variance of the folded β statistic

The formulation for $\beta^{(1)*}$ does not fall into the class of neutrality tests based on the folded site frequency spectrum studied in Achaz (2009), because the folded frequency of each SNP is not considered in our formulation. Therefore, we provide a derivation below. ϕ and ρ are defined in Achaz (2009) and d_i is the measure of frequency similarity. $S_g(i)$ is the number of SNPs in the window of folded frequency g(i) and is analogous to η_i in Achaz (2009). Set $m = \lceil \frac{n}{2} \rceil$, where $\lceil \rceil$ denotes the ceiling. We refer to the estimator of $\hat{\theta}_\beta^{fold}$ reported in Siewert and Voight (2017) as $\hat{\theta}_\beta^*$. The variance

of $\beta^{(1)*}$ is:

$$Var[\hat{\theta}^*_\beta - \hat{\theta}_W] = Var[\hat{\theta}^*_\beta] + Var[\hat{\theta}_W] - 2Cov[\hat{\theta}^*_\beta, \hat{\theta}_W] \tag{2.3.2}$$

First, we derive the variance of $\hat{\theta}^*_\beta$ to be:

$$\begin{aligned}
Var[\hat{\theta}^*_\beta] &= Var\left[\frac{\sum_{i=1}^m d_i S_{g(i)}}{\sum_{i=1}^m d_i(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}}\right] \\
&= \left(\sum_{i=1}^m d_i\Big(\frac{1}{i}+\frac{1}{n-i}\Big)\frac{1}{1+\delta_{i,n-i}}\right)^{-2} Var\left[\sum_{i=1}^m d_i S_{g(i)}\right] \\
&= \left(\sum_{i=1}^m d_i\Big(\frac{1}{i}+\frac{1}{n-i}\Big)\frac{1}{1+\delta_{i,n-i}}\right)^{-2}\left(\sum_{i=1}^m Var[d_i S_{g(i)}] + \sum_{i\neq j} Cov[d_i S_{g(i)} d_j S_{g(j)}]\right) \\
&= \left(\sum_{i=1}^m d_i\Big(\frac{1}{i}+\frac{1}{n-i}\Big)\frac{1}{1+\delta_{i,n-i}}\right)^{-2}\left(\sum_{i=1}^m d_i^2(\phi_i\theta+\rho_{ii}\theta^2) + \sum_{i\neq j} d_i d_j \rho_{ij}\theta^2\right) \\
&= \left(\sum_{i=1}^m d_i\Big(\frac{1}{i}+\frac{1}{n-i}\Big)\frac{1}{1+\delta_{i,n-i}}\right)^{-2}\left(\sum_{i=1}^m d_i^2(\phi_i\theta+\rho_{ii}\theta^2) + 2\sum_{1\leq i\leq j\leq m} d_i d_j \rho_{ij}\theta^2\right)
\end{aligned} \tag{2.3.3}$$

Next, the $Var[\hat{\theta}_W]$ is taken from Achaz (2009):

$$\begin{aligned}
Var[\hat{\theta}_W] = \Big(\sum_{i=1}^m \frac{n}{i(n-i)(1+\delta_{i,n-i})}\Big)^{-2}\Bigg(&\theta\Big(\sum_{i=1}^m \Big(\frac{n}{i(n-i)(1+\delta_{i,n-i})}\Big)^2 \phi_i^{-1}\Big) \\
&+ \theta^2\Big(\sum_{i=1}^m \Big(\frac{n}{i(n-i)(1+\delta_{i,n-i})}\Big)^2 \phi_i^{-2}\rho_{ii} \\
&+ 2\sum_{i=1}^m\sum_{j=i+1}^m \phi_i^{-1}\phi_j^{-1}\frac{n}{i(n-i)(1+\delta_{i,n-i})}\frac{n}{j(n-j)(1+\delta_{j,n-j})}\rho_{ij}\Big)\Bigg)
\end{aligned} \tag{2.3.4}$$

Finally, the covariance of $\hat{\theta}^*_\beta$ and $\hat{\theta}_W$:

$$\begin{aligned}
Cov[\hat{\theta}^*_\beta, \hat{\theta}_W] &= Cov\left[\frac{\sum_{i=1}^m d_i S_{g(i)}}{\sum_{i=1}^m d_i(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}}, \frac{\sum_{i=1}^m S_{g(i)}}{\sum_{i=1}^m(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}}\right] \\
&= \frac{1}{\sum_{i=1}^m d_i(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}}\frac{1}{\sum_{i=1}^m(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}} Cov\Big[\sum_{i=1}^m d_i S_{g(i)}, \sum_{i=1}^m S_{g(i)}\Big] \\
&= \frac{1}{\sum_{i=1}^m d_i(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}}\frac{1}{\sum_{i=1}^m(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}} \sum_{i=1}^m\sum_{j=1}^m d_i Cov[S_{g(i)}, S_{g(i)}] \\
&= \frac{1}{\sum_{i=1}^m d_i(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}}\frac{1}{\sum_{i=1}^m(\frac{1}{i}+\frac{1}{n-i})\frac{1}{1+\delta_{i,n-i}}} \sum_{i=1}^m\sum_{j=1}^m d_i \rho_{ij}\theta^2
\end{aligned} \tag{2.3.5}$$

2.3.3 Standardized β statistics

The standardized $\beta^{(1)}$ statistics are given by:

$$\beta^{(1)}_{std} = \frac{\beta^{(1)}}{\sqrt{Var[\beta^{(1)}]}} = \frac{\hat{\theta}_\beta - \hat{\theta}_W}{\sqrt{\alpha^*_n\hat{\theta} + \beta^*_n\hat{\theta}^2}} \tag{2.3.6}$$

$$\beta^{(1)*}_{std} = \frac{\beta^{(1)*}}{\sqrt{Var[\beta^{(1)*}]}} = \frac{\hat{\theta}^*_\beta - \hat{\theta}_W}{\sqrt{Var[\hat{\theta}^*_\beta] + Var[\hat{\theta}_W] - 2Cov[\hat{\theta}^*_\beta, \hat{\theta}_W]}} \tag{2.3.7}$$

2.4 Window size containing signature of balancing selection

Although β can be calculated on any window size, previous work has suggested that the effects of balancing selection are localized to a narrow region surrounding the balanced site (Gao *et al.*, 2014). Ultimately, the optimal window size depends on the recombination rate, as it breaks up alleleic classes.

If one uses too small of window size, then some of the signal of alleleic-class build up will be excluded from the statistic, reducing power. It too large of window size is used, then noise from regions beyond those which provide any signal of selection will decrease power. Optimally, we could calculate β on the window which has not experienced much, if any, recombination between alleleic classes. According to our model, this region will contain all variants fixed in alleleic class, and potentially some variants which were once fixed in alleleic class but have drifted slightly in frequency due to recombination beginning to decouple them from selection.

The probability of recombination between allelic classes is equal to the total coalescent branch length in the allelic class multiplied by the probability of recombination onto the other allelic class. Because we are detecting long-term selection, most of the coalescent branch length will fall into the portion between coalescence within each allelic class and coalescence of the two allelic classes. We can, therefore, put an upper bound on the size of the ancestral region. The probability of any recombination event occurring at a certain position at any time point in T generations is ρT, where ρ is the

individual recombination rate. The probability of a recombination event occurring between a chromosome from allelic class 1 and any chromosome from allelic class 2, given that a recombination event occurs in a chromosome from class 1, is just the frequency of allelic class 2. Similarly, the probability that if a recombination event occurs in class 2, it is with any chromosome from class 1, is just the frequency of allelic class 1. Let λ be the rate of observable recombination, in units of base pairs, where p and q are the frequencies of the 2 allelic classes, which must sum to 1 by definition.

$$\lambda = T\rho p + T\rho q$$

$$\lambda = T\rho$$

The distribution of the length of the ancestral segments on either side of the balanced loci is then exponential with rate parameter $T\rho$.

For our analysis of the 1000 Genomes Project, we are focusing on detecting events that occurred after a split with chimpanzee, but that are old enough that our method has power. Assuming a recombination rate of 2.5×10^{-8} per individual per basepair and a split time of $250,000$ generations prior with selection starting at this same time, the 95th quantile on either side is then 479 basepairs. The most recent events we can hope to detect are closer to $100,000$ generations prior to present, giving a 95th quantile of 1198 bases on either side of the core SNP. Based on these estimates, we chose to perform our analysis using a window size of 500 base pairs on either side of

the core site, for a total size of 1kb. This matches the window size with optimal power for each summary statistic in simulations with the recombination rate 2.5×10^{-8} **(Fig. 2.8)**.

2.5 Power analysis

2.5.1 Simulations

We generated two sets of simulations: one without a balanced variant (the set we refer to as our neutral simulations) and one with a balanced variant (balanced simulations) using the forward genetic simulator SLiM 2 (Haller and Messer, 2017). In the second set, a single balanced variant was introduced at the center of the simulated region in the human population, either at the time of speciation (250,000 generations prior to simulation ending), or 150,000 generations after speciation (100,000 generations prior to simulation ending). The simulations then continued as normal, conditional on maintenance of the balanced SNP in the population. If this balanced variant was lost, the simulation restarted at the generation in which the balanced variant was introduced. In the second (neutral) set, no balanced variant was introduced, so all variants are selectively neutral.

Each balanced SNP had an overdominance coefficient h and selection coefficient s. The fitness of the heterozygote is then 1+hs, and the fitness of the ancestral and

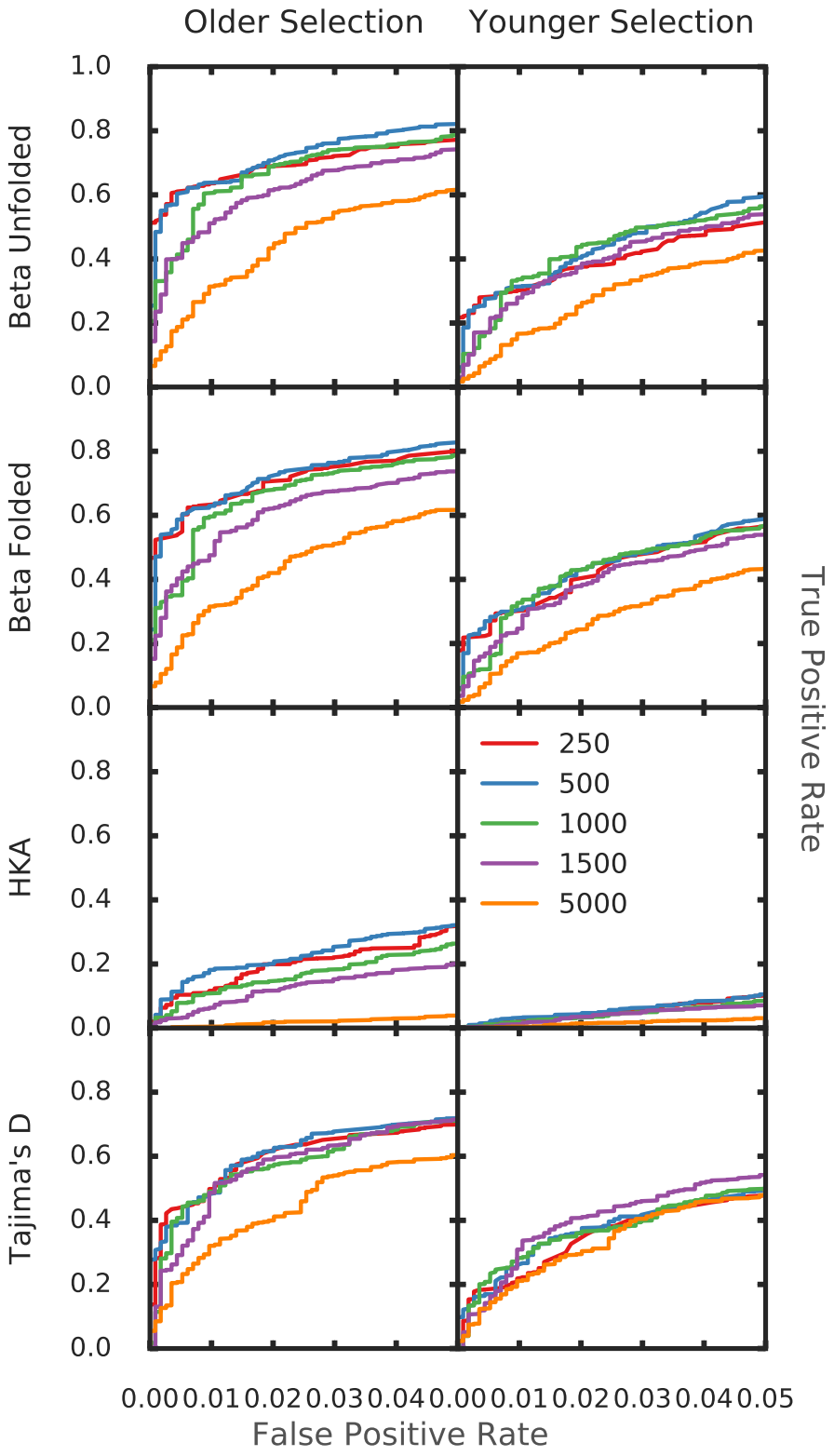

Figure 2.8: Power to detect ancient balancing selection using different window sizes, in units of base pairs, with an equilibrium frequency of 0.5.

derived homozygotes are 1 and 1+s, respectively. We simulated two different s values: 10^{-2} (our default) and 10^{-4}. We simulated six different equilibrium frequencies: 0.17, 0.25, 0.5, 0.75, 0.83, which correspond to h=0.25, 0.5, 100, 1.5, 1.25. Negative h values were paired with negative s values.

2.5.2 Method of power comparison

After simulation completion, the frequency of each variant in the sampled individuals was calculated. Our default sample size was 100 haploid individuals. Substitutions were defined as any variant in which the allele from the chimpanzee chromosome was not found in the sampled human individuals. For each set of balanced simulations, we define the core SNP as the variant under balancing selection. For each set of balanced simulations, we then found a corresponding set of core SNPs in our neutral simulations which were within 10% of the equilibrium frequency of the balanced variants. We then calculated the score for each statistic on these core variants. In this way, we have statistic scores for the balanced variant from each balanced simulation replicate, and a score for a neutral variant matched for similar frequency.

To calculate the power of each method, we compared the score of the balanced variant in balanced simulations with the score of SNPs matched for equilibrium frequency in neutral simulations. For each neutral simulation replicate, we randomly identified one SNP in the simulated region at a frequency within 10 percent of the equilibrium frequency of the corresponding simulations with a balanced SNP. Throughout our

discussion of simulations, we refer to the number of the haploid genotypes, corresponding to the total number of chromosomes, as the number of individuals. Power calculations were performed with p=20 for β.

For $T1$ and T2, a number of informative sites of about 20, or 10 on either side of the core site, achieved maximum power in simulations **(Fig. 2.9)**. Furthermore, this roughly matches the expected number of informative sites in a 1-kb region under selection. Therefore, a window of 20 total informative sites is roughly equal to the expected ancestral region size, which is roughly equal to the window at which all these methods achieve optimal power. For this reason, we used a 1-kb window or 20 informative sites, as applicable, when calculating each statistic.

The $T1$ and $T2$ statistics require an estimate of divergence time with the outgroup species and a summary of the background levels of polymorphisms and substitutions. To generate these empirical genome-wide estimates, we pooled all of our neutral simulation replicates for the appropriate parameter set, and then inputted these into the functions provided by BALLET, the software package implementing the $T1$ and $T2$ statistics (DeGiorgio *et al.*, 2014).

To generate expectation and variance for the HKA test, we took 1kb regions from each of our neutral simulations under the relevant parameter set. We then calculated the mean and variance of the number of sites that are polymorphic in the human simulated population, and of the average number of differences between a random human individual and the chimp outgroup individual. Our HKA statistic was then

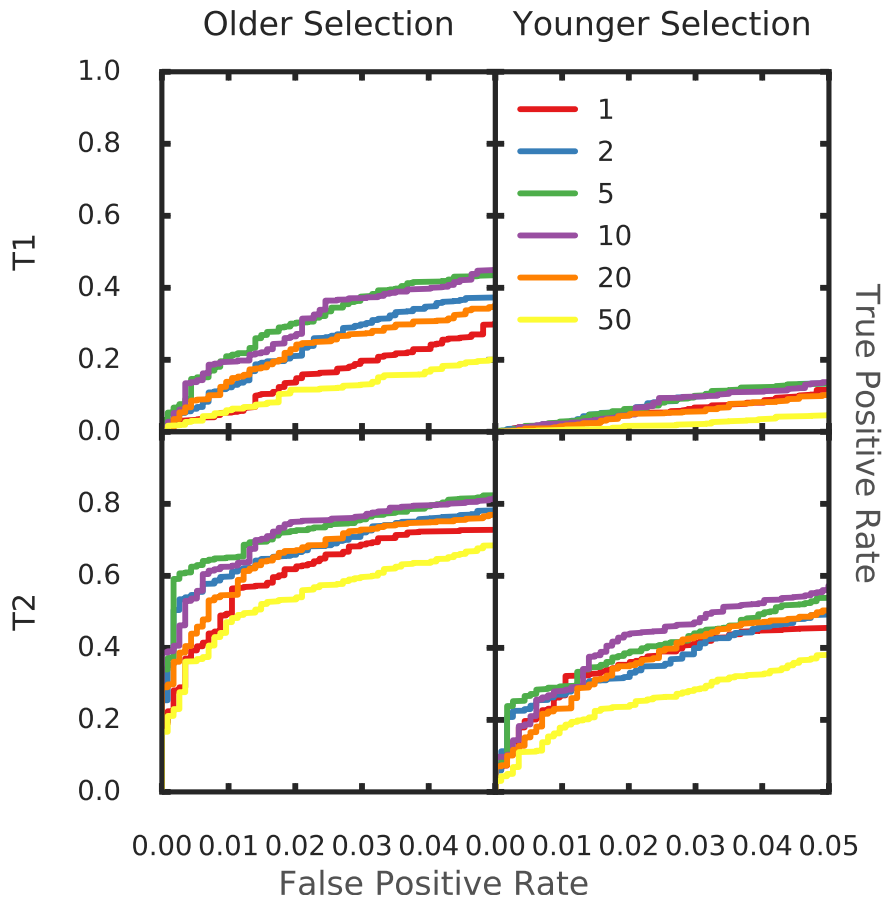

Figure 2.9: Power to detect ancient balancing selection using different numbers of informative sites. The number of sites corresponds to the number of sites on either side of the core site.

the sum of two chi-squared statistics: one corresponding to the number of human polymorphisms, and one corresponding to the average number of human/chimp differences.

For the mutation and recombination rate variation power analysis, we used the background files generated using the simulations based on our default rates. The reason for this is to both check for over-fitting to these parameters and also to test for power upon misspecification of population parameters.

2.5.3 Power comparison results

Compared to other summaries, β had the greatest performance across most parameter combinations **(Fig. 2.10)**. As expected, the $\beta^{(1)}$ statistics performs slightly worse than $T2$ under many conditions. However, unlike T2, our method does not require an outgroup sequence or grids of simulations which are computationally expensive.

We next investigated the power of β under more complex demographic scenarios representative of recent human history (DeGiorgio *et al.*, 2014). We found that β performs well under bottleneck and expansion models. Under an expansion scenario, the performance of all methods decreased **(Fig. 2.11)**, consistent with results from previous studies (DeGiorgio *et al.*, 2014), possibly due to the larger population size increasing the expected time until an allele can fix in its allelic class. The effect of a population bottleneck on power was less drastic and led to a slight increase in power

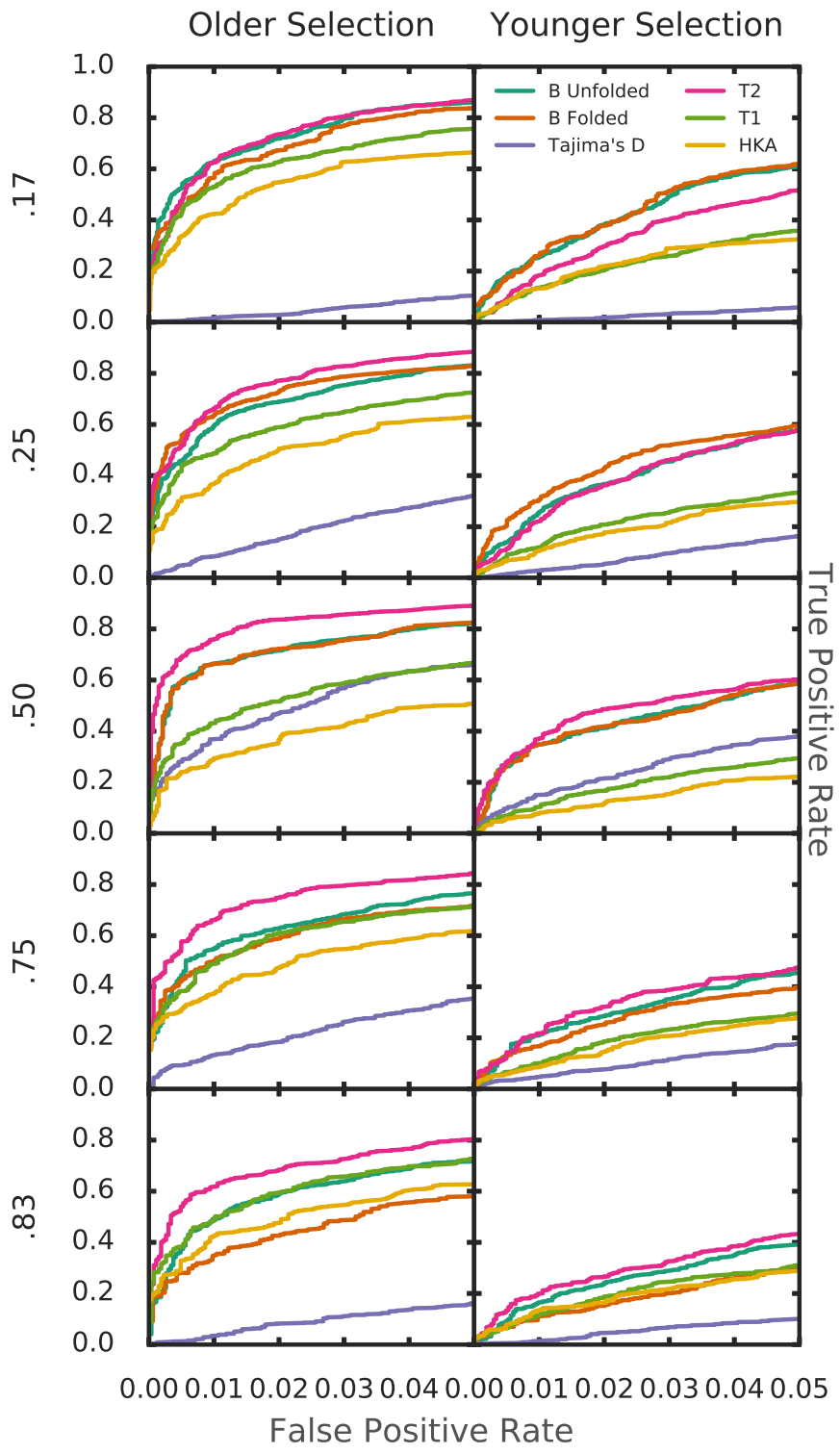

Figure 2.10: Power to detect ancient balancing selection under equilibrium demography. Rows correspond to different equilibrium frequencies.

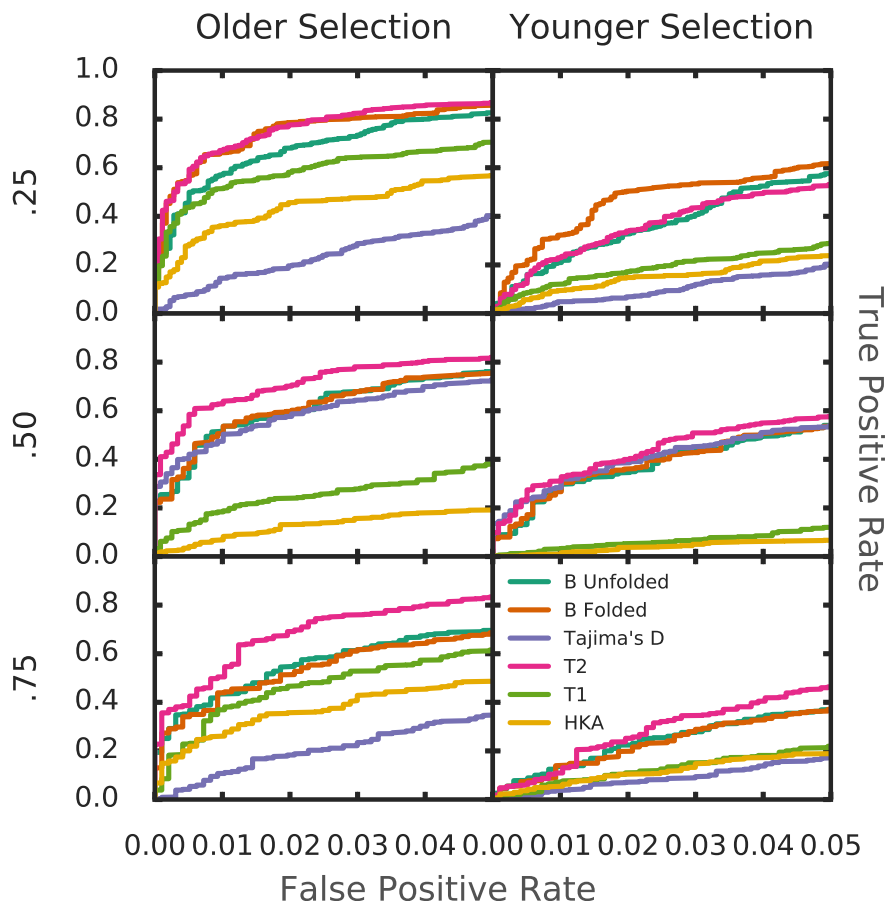

Figure 2.11: Power of methods to detect ancient balancing selection under a model of population expansion. In this demographic model, the human population expands to $N_e = 20,000$ at generation 302,000, then remains that size until sampling. Based on rescaled simulations.

to detect more recent selection **(Fig. 2.12)**.

Population substructure can confound scans for selection (Ingvarsson, 2004; Schierup *et al.*, 2000). To investigate the power of our method in these scenarios, we simulated two models of population substructure. First, we considered a model of two completely subdivided populations. We pooled together 50 individuals from each subpopulation with which to perform the statistical calculations. In this case, the power of all methods to detect balancing selection at equilibrium frequency 0.5 decreased

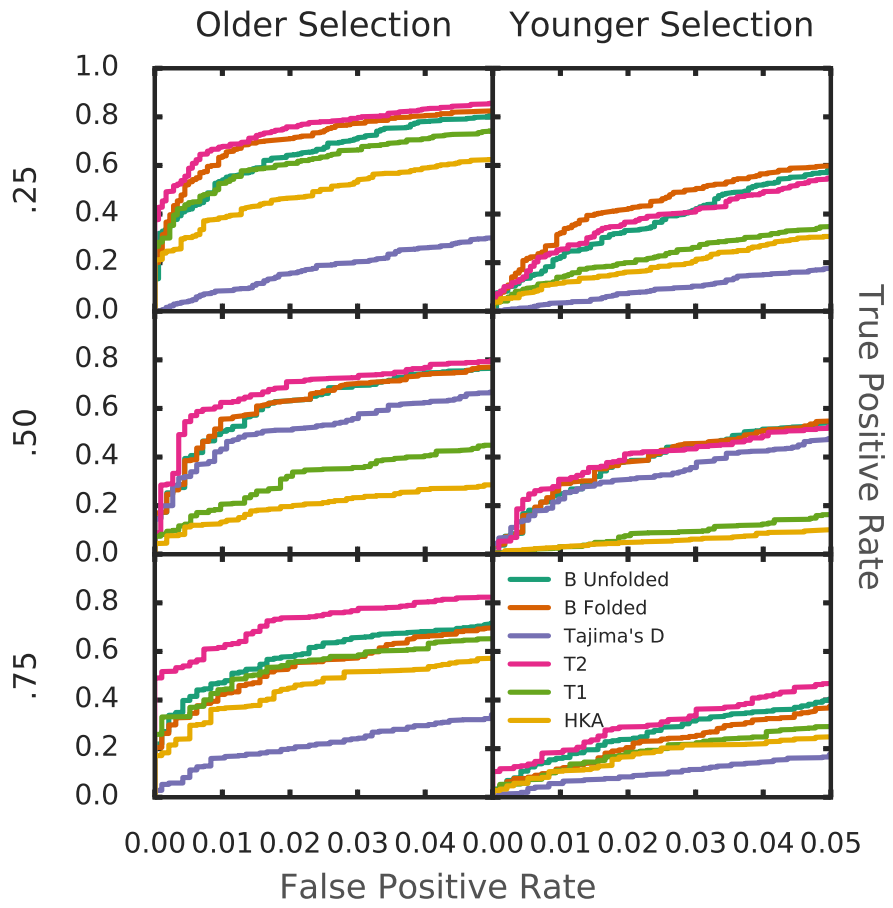

Figure 2.12: Power of methods to detect ancient balancing selection under a model of a population bottleneck. Based on rescaled simulations. In this scenario, human population size drops to $N_e = 5,500$ from generations 320,000 to 328,000, then returns to $N_e = 10,000$. Based on rescaled simulations.

considerably **(Fig. 2.13)**. This matches expectation, as this situation is expected to drastically increase the number of variants at frequency 0.5.

Next, we considered a two-pulse model of ancient admixture. We selected this model because of its approximation of Neanderthal admixture into human (Vernot and Akey, 2015), which may be thought to confound scans for selection in humans. Power with Neanderthal admixture stayed roughly the same as without **(Fig. 2.14)**. This is as expected, as most haplotypes introduced through admixture are expected to be at very low frequency so will not reach the frequency of the balanced SNPs or matched neutral SNPs.

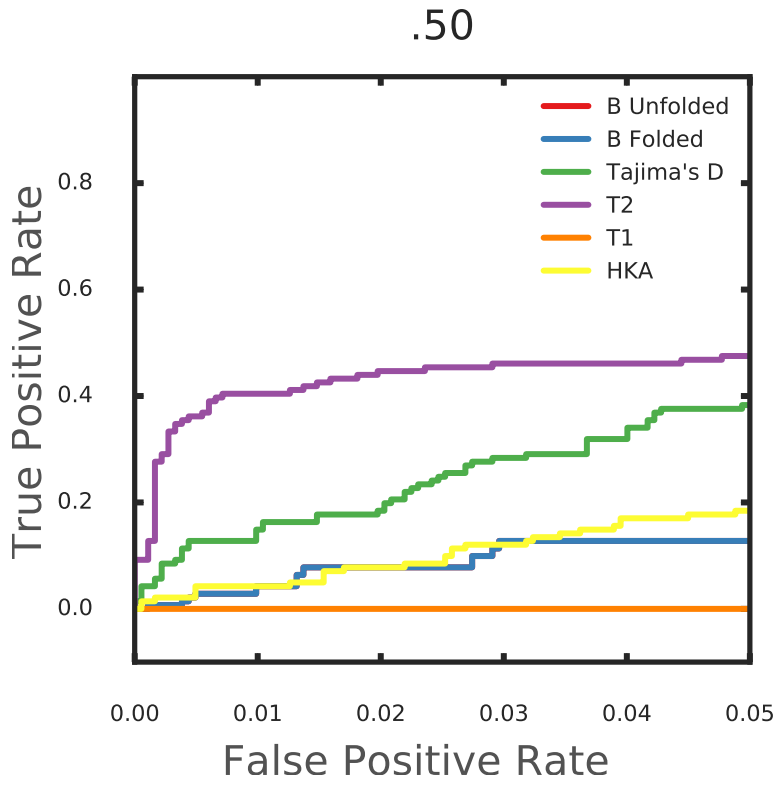

Figure 2.13: Power of methods to detect ancient balancing selection under a model of complete population subdivision. In this case, the human population is completely divided into two subpopulations of equal size, $N_e = 5000$, at generation number 300,000, with no admixture between them. The subpopulations were then combined to calculate allele frequencies. This represents an extreme case: there are expected to be a large number of variants at frequency 0.5. In this analysis, we excluded simulation replicates in which the core SNP was not of frequency exactly 0.5, in order to investigate the power at the exact frequency that variants are expected to accrue due to population substructure. Balanced variants were of the "older selection" category, so were introduced at generation 100,000. For this analysis, we used the empirical background files from the corresponding neutral simulations, but the estimated divergence time from the simulations using our default rates. This is because the simple divergence time estimator included in BALLET is not able to accurately infer divergence times with the outgroup in the presence of significant population structure. We note that Beta Folded and Unfolded perform nearly identical in this case. Based on rescaled simulations.

We next examined the power for all methods under models of variable mutation rates, recombination rates, and sample sizes. As expected, the power of all methods was positively correlated with mutation rate **(Fig. 2.15, 2.16)**, and negatively correlated with recombination rate **(Fig. 2.17, 2.18)**. A higher mutation rate provides more variants that can accumulate within an allelic class, whereas a lower recombination rate causes longer haplotypes upon which mutations can accumulate.

β has reasonable power down to very small sample sizes, achieving near maximum power with as few as 20 sampled chromosomes **(Fig. 2.19, 2.20)**. In practice, the sample size used to calculate the frequency of each variant may differ between variants. We tested the power of β when the sample size of each variant is downsampled from the original size of 100 by a random amount from 0 to 25 individuals. We found that this decreases power very slightly, and that lower values of p perform better in this scenario **(Fig. 2.21)**.

Finally, power remained high under frequency-dependent selection **(Fig. 2.22)**, and when a lower selection coefficient was simulated **(Fig. 2.23)**. This matches expectation, as frequency-dependent selection is expected to maintain haplotypes in the population for long time periods, causing allelic class build-up. A lower selective coefficient would be expected to lower the probability of maintenance of the balanced allele in the population, but conditioned on this maintenance, should not affect power, as we observed.

Simulations show that the power of the folded version of β is similar to the unfolded

version at intermediate allele frequencies, but has reduced power at very high frequencies **(Fig. 2.10)**. However, even at these frequencies, it still outperforms Tajima's D, the only other statistic of those tested which does not require knowledge of the ancestral state or an outgroup.

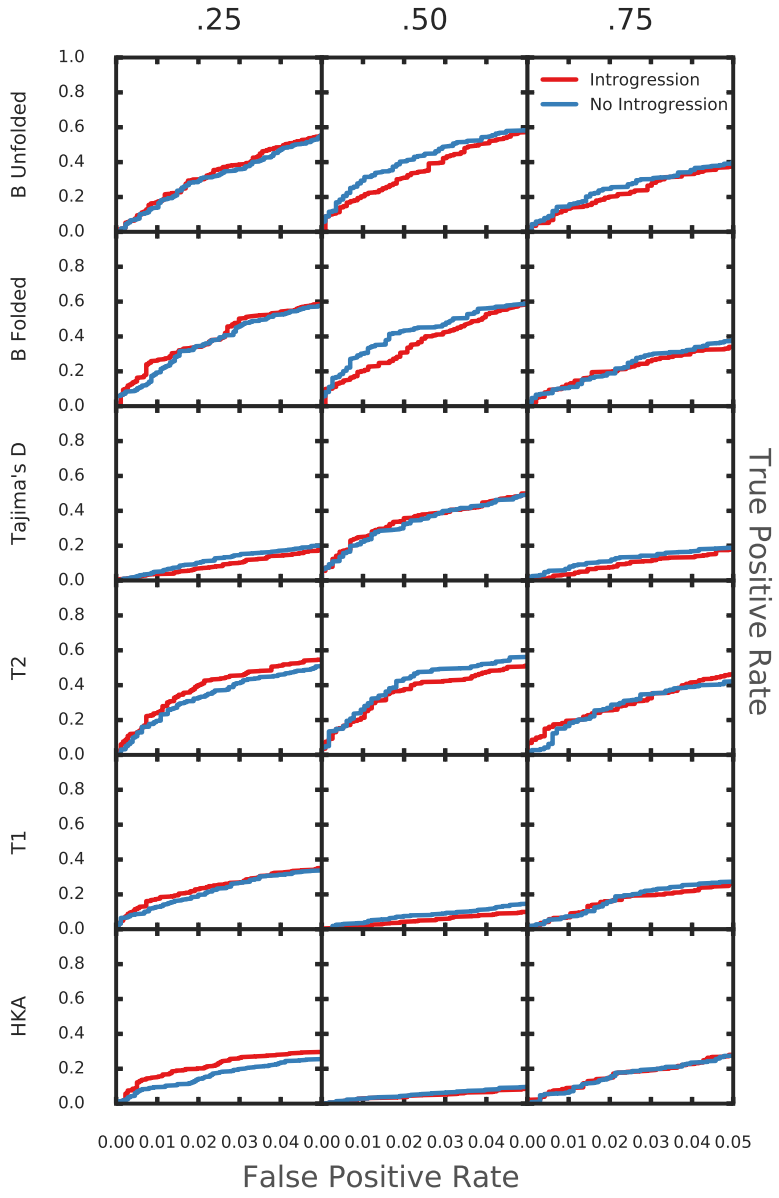

Figure 2.14: Power to detect ancient balancing selection under admixture. Balanced variants were introduced at generation 250,000. In this scenario, we simulated Neanderthal admixture into Asian populations. Based on the demographic model presented in Vernot and Akey (2015), we used a two pulse model, with a split from the human lineage into Neanderthal at generation $315,000$ with an N_e of 1500. The first pulse of Neanderthal admixture into human occurred from generations $347,750$ to $347,780$, with a migration rate of .00075. The second, weaker pulse occurred from generation $347,820$ to $347,850$, with migration rate .0002. The human N_e and chimpanzee remained our default.

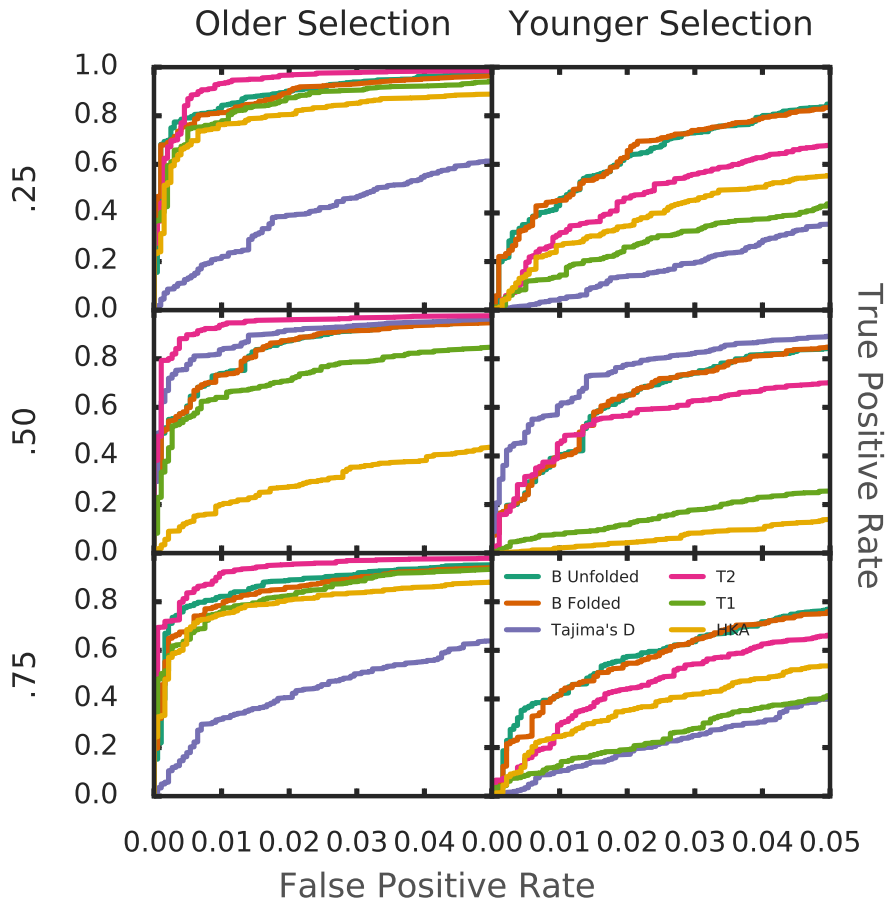

Figure 2.15: Power of methods with an increased mutation rate of 2.5×10^{-7}.

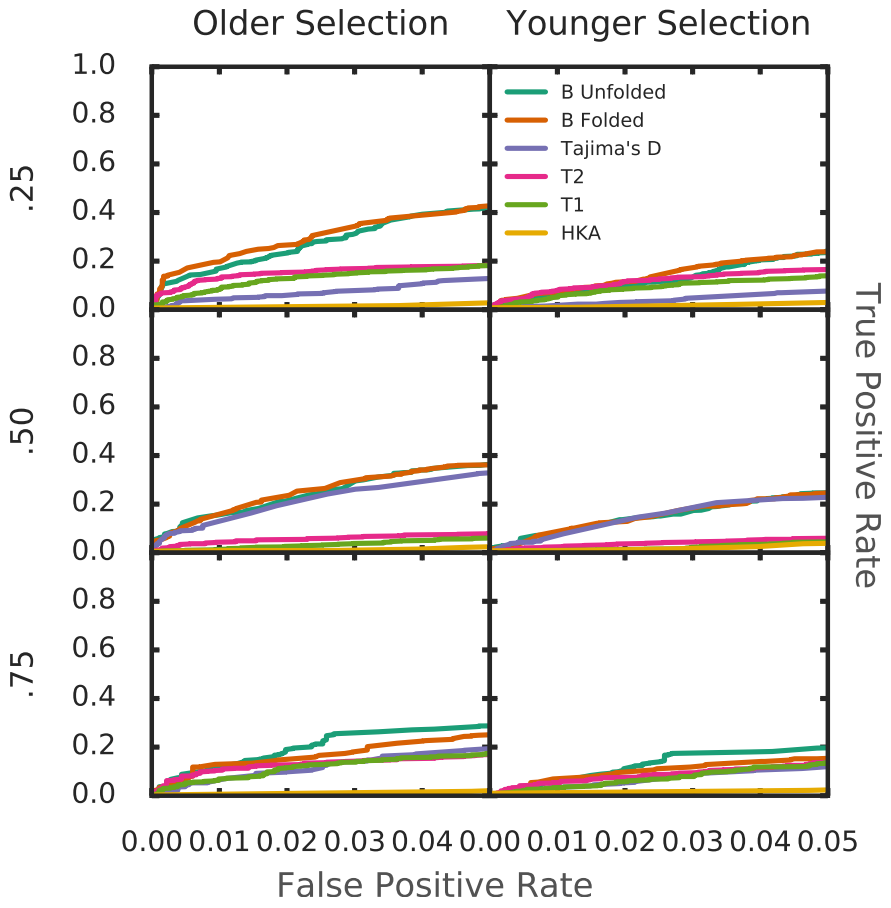

Figure 2.16: Power of methods with a decreased mutation rate of 2.5×10^{-9}. We note that $T1$ and $T2$ perform poorly due to there not being 20 informative sites in the 10 kb simulated region, which results in an error.

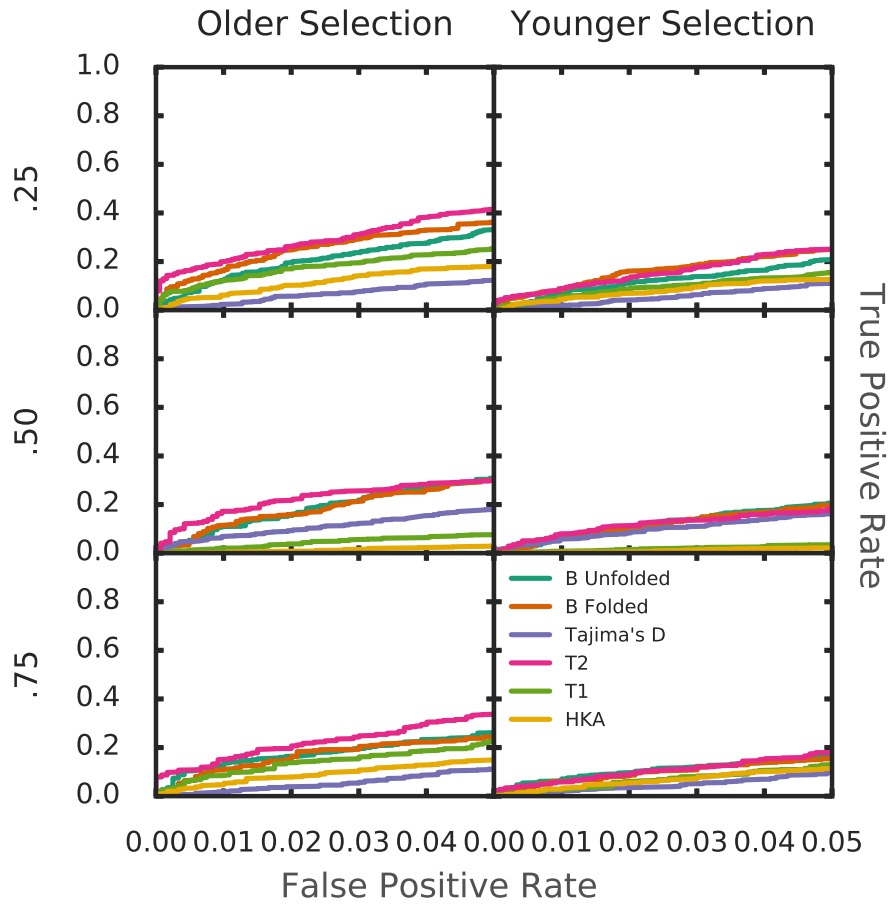

Figure 2.17: Power of methods with an increased recombination rate of 2.5×10^{-7}.

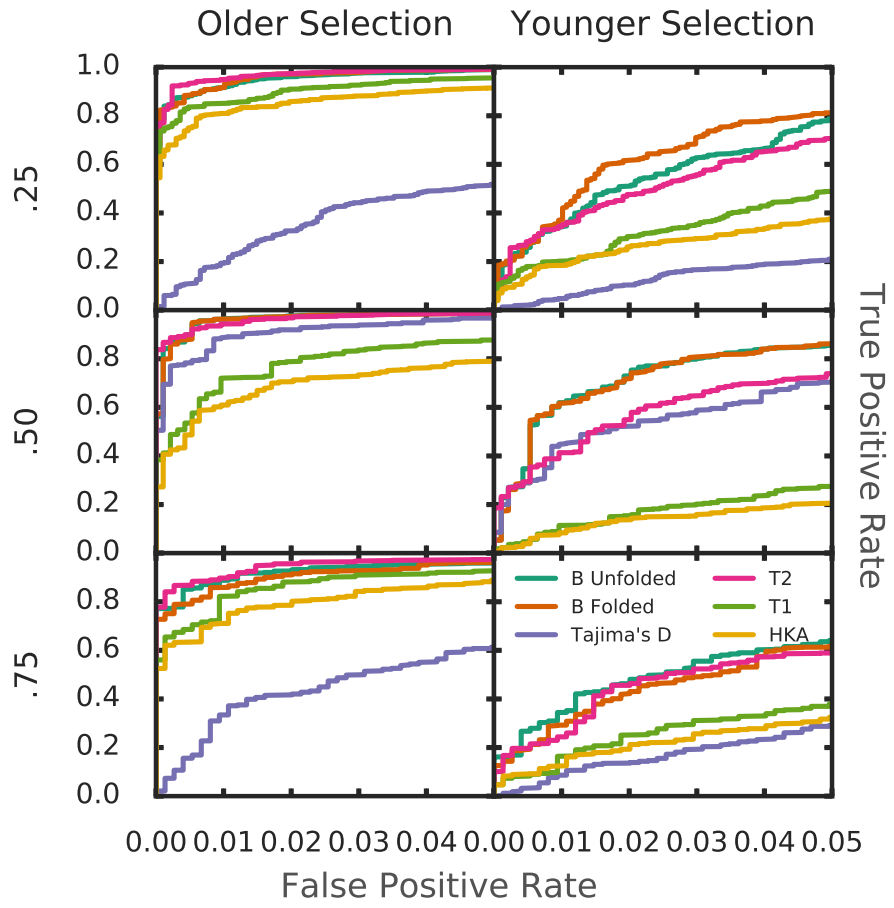

Figure 2.18: Power of methods with a decreased recombination rate of 2.5×10^{-9}.

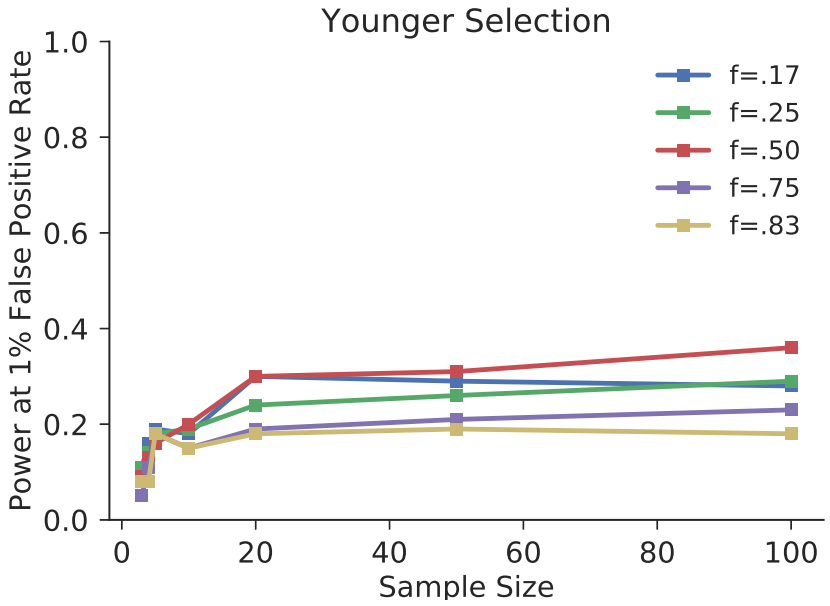

Figure 2.19: Power of β at a 1 percent false positive rate to detect selection 100,000 generations old, by number of chromosomes sampled and at different equilibrium frequencies (f).

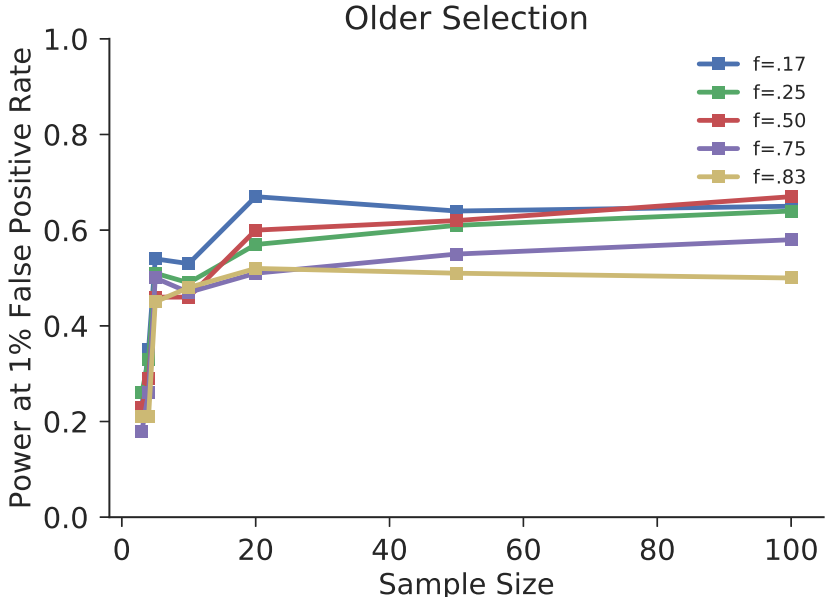

Figure 2.20: Power of β at a 1 percent false positive rate to detect selection 250,000 generations old, by number of chromosomes sampled and at different equilibrium frequencies (f).

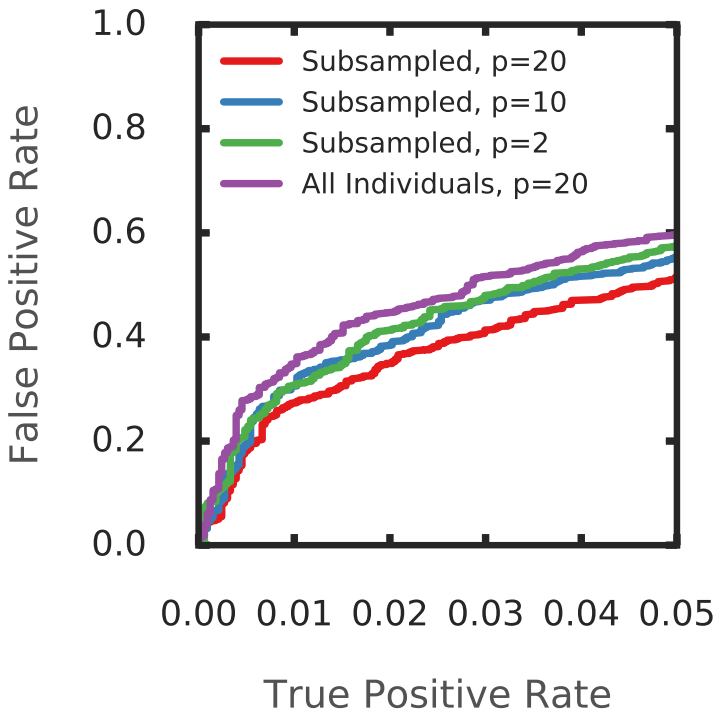

Figure 2.21: Power of $\beta^{(1)}$ to detect balancing selection when SNP frequencies are calculated using different numbers of individuals and different values of the p parameter. In order to investigate the effects this has on power, we subsampled individuals from our initial set of 100. For each SNP in each simulation replicate, we chose a number uniformly, between 0 and 25, of individuals to remove. After these individuals were removed the frequency was recalculated on the remaining individuals.

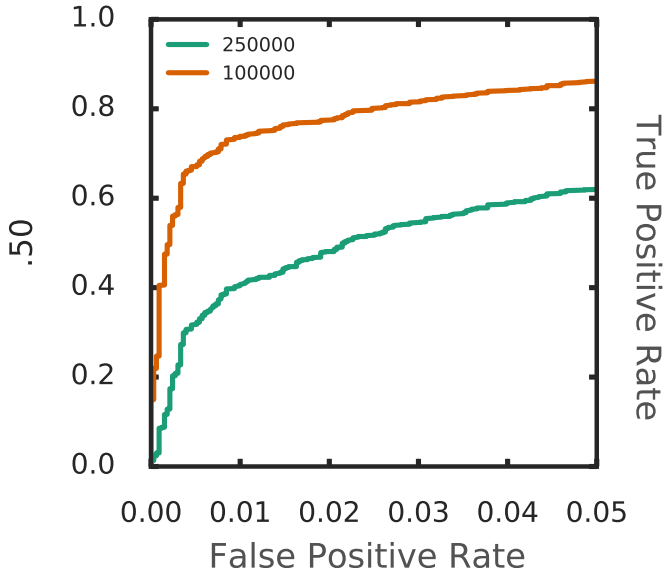

Figure 2.22: Power of the $\beta^{(1)}$ statistic under a model of frequency-dependent selection. In this case, the fitness coefficient was .01 and the overdominance coefficient was .05. The fitness of the derived allele was set to equal 1.5 minus the frequency of the allele. This results in an equilibrium frequency of .5. The color corresponds to age of selection, either 100,000 generations after the start of selection (older selection) or 250,000 generations after the start(younger selection).

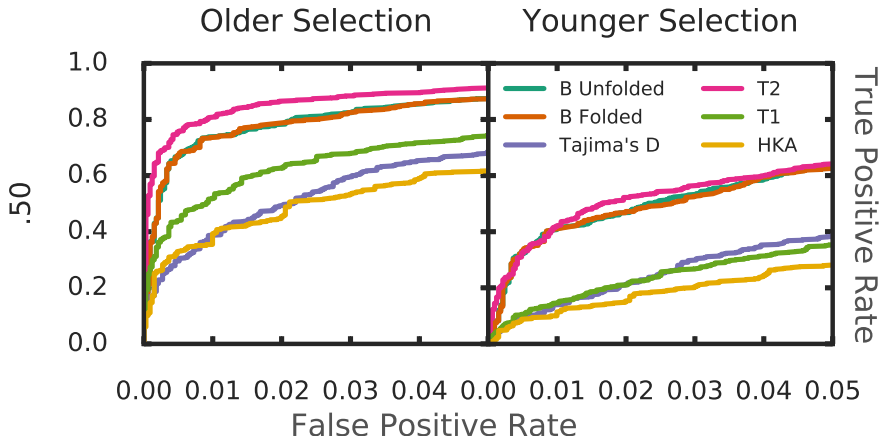

Figure 2.23: Power of methods with a selective coefficient of 1×10^{-4} and overdominance coefficient of $h = 100$. We were only able to test power with $h = 100$, because of the extremely high frequency at which the balanced allele was lost at other equilibrium frequencies.

Chapter 3

Application of $\beta^{(1)}$ to detect balancing selection in humans

3.1 Overview of scan

We applied $\beta^{(1)}$ to population data obtained by the 1000 Genomes Project (Phase 3) to detect signatures of balancing selection (The 1000 Genomes Consortium, 2015). We calculated the value of β in 1kb windows around each SNP in all 26 populations, separately. We focused on regions that passed sequencing accessibility and repeat filters **(Section 3.2)**. β scores appeared well-calibrated and consistent across populations **(Fig. 3.1)**.

We defined extreme β scores as those in the top 1% in the population under consideration **(Section 3.2)**. We analyzed the autosomes and X-chromosome separately.

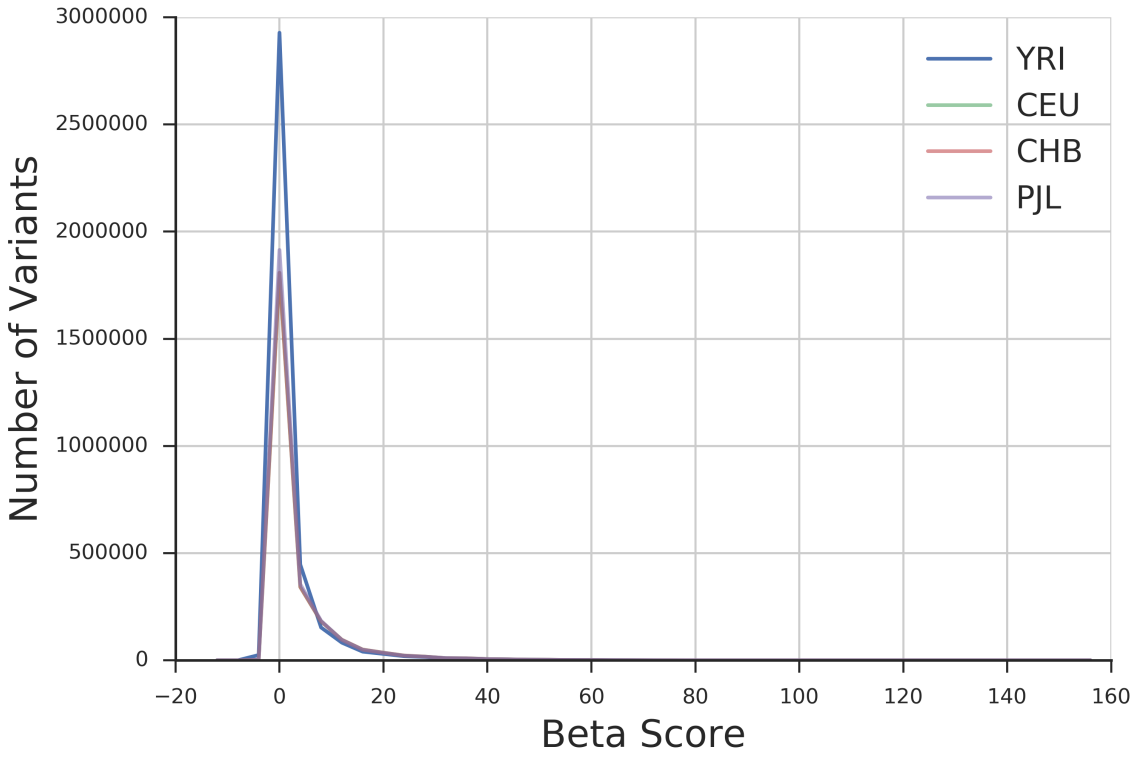

Figure 3.1: Distribution of Beta in 4 representative populations. Beta scores binned in units of 4.

Because our method is designed to detect ancient balancing selection, we focus on signals of selection that predate the split of modern populations. For this reason, we further filtered for loci that were top-scoring in at least half of the populations tested. We focus on results of our unfolded β scan, however, we also scanned using the folded β statistic to test for robustness of our top scoring sites.

We identified 8,702 autosomal, and 317 X-chromosomal, top-scoring variants that were shared among at least half (≥ 13) of the 1000 Genomes populations. Together, these variants comprise 2,453 distinct autosomal and 86 X-chromosomal loci, and these signatures overlapped 692 autosomal and 29 X-chromosomal genes.

3.2 Methods for 1000 Genomes Analysis

To apply our method to 1000 Genomes data, we first downloaded data for each of the 26 populations in phase 3 of the project (obtained May 2nd, 2013). We then calculated allele frequencies separately for each population, and calculated β in 1 kb sized windows centered around each SNP for each population. We filtered out variants which did not have a folded frequency of at least 15% in a minimum of one population. The purpose of the frequency filter is to prevent false positives: we were unable to simulate balancing selection with a folded equilibrium frequency of less than 15%, due to the high probability of one allele drifting out of the population, as expected due to theory **(Section 1.2.1)**. Therefore, variants with a high β score but a folded frequency less than 15% have a high likelihood of being false positives.

Because poorly sequenced regions can artificially inflate the number of SNPs in a region, we then filtered out regions that contained one or more base pairs that were ruled as poor quality in the 1000 Genomes phase 3 strict mask file. For further confirmation that the signal was not a result of poor mapping quality, we overlapped SNPs of interest with hg19 human RepeatMasker regions, downloaded from the UCSC Table Browser on February 9th, 2017. We then removed all core SNPs from consideration that were found within a repeat, similar to Bubb *et al.* (2006). We further removed SNPs that were not of common frequency (at or above a folded frequency of 15%) in at least one population. After filtering, there were 1,803,299 SNPs that remained. We then found the top 1% of these high-quality SNPs in each population in our β

scan.

Unknown paralogs or other technical artifacts could inflate the number of intermediate frequency alleles. Although the 1000 Genomes data provides strict quality filter masks, we wanted to further verify that our haplotypes of interest in *WFS1* and *CADM2* were not the result of obvious technical artifacts. In order to do this, we used the –hardy flag in vcftools (Danecek *et al.*, 2011), and investigated both the one-tailed p-value for an excess of heterozygotes, and the two-tailed p-value, in our 4 representative populations (YRI: Yoruban from Africa, CEU: Utah Residents with Northern and Western European Ancestry, CDX: Chinese Dai, and PJL: Punjabi). All variants on these haplotypes had p-values above 1×10^{-3}.

The lowest autosomal significance cut-off of any population, ASW, corresponds to a β score of 47.49. This score is in the top 0.05 percentile of core SNPs in neutral simulations corresponding to an equilibrium frequency of 0.5 **(Fig. S3)**.

To find top-scoring sites that are also GWAS hits, we obtained LD proxies in European populations for our top-scoring SNPs, using a cut-off of r^2 of 0.9, a maximum distance of 50kb and a minimum minor allele frequency of 5%. We then overlapped these LD proxies with GWAS hits obtained from the GWAS Catalog to get our final list of putatively balanced GWAS hits (Welter *et al.*, 2014) **(Table 3.1)**. Gene names and locations were downloaded from Ensembl BioMart on November 26th, 2016.

For our trSNP comparison, we used the Human/Chimp shared haplotypes from Leffler *et al.* (2013). Using logistic regression, we then modeled the outcome of a SNP being

part of a trHap as dependent on the β Score and distance to nearest gene.

3.3 Characterization of signals

Trans-species haplotypes are highly unlikely to occur by chance, unlike trans-species SNPs, which are expected to be observed in the genome due to recurrent mutations (Gao *et al.*, 2014). These haplotypes present a signature of balancing selection independent from the signature captured by β. If β captures true signatures of balancing selection, one would expect an enrichment of high β values at trans-species haplotypes. We found that β was in fact predictive of trans-species haplotype status from Leffler *et al.* (2013), even after including adjustments for the distance to the nearest gene ($P < 2 \times 10^{-16}$) **(Section 3.2)**. However, out of 125 trans-species haplotypes from Leffler *et al.* (2013) only 6 are in the top percentile in the Yoruban (YRI) population. Although this represents an enrichment, it is perhaps lower than one would expect. We hypothesize that this is due to the lower power of β in regions with a lower effective mutation rate, as would be expected in regions of high selective constraint. These regions of higher selective constraint are enriched for trans-species haplotypes (Leffler *et al.*, 2013), confounding this analysis.

Our scan identified several loci that have been previously implicated as putative targets of balancing selection. Several major signals occurred on chromosome 6 near the HLA, a region long presumed to be subjected to balancing selection Hedrick (1998);

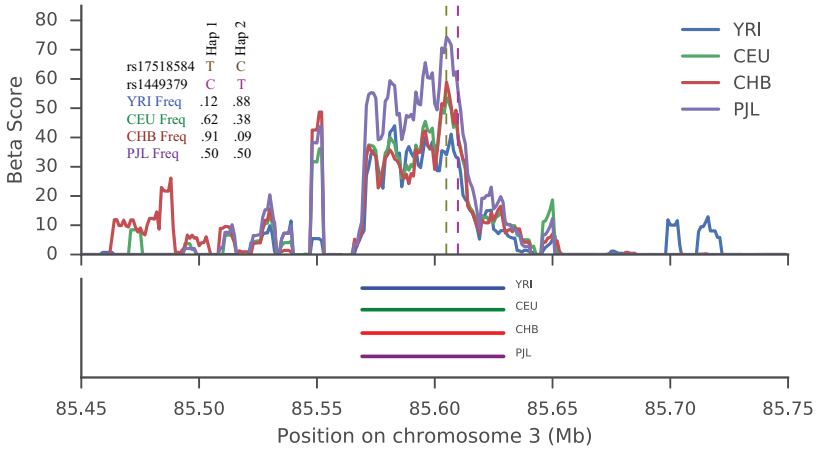

Figure 3.2: Signal of balancing selection at *CADM2*. The signal of selection is located in an intron of *CADM2*. (a) rs17518584 is the lead GWAS SNP for several intellectual traits and is marked by the brown vertical dashed line. The purple dashed line marks two regulatory variants found on the balanced haplotype. β scores were calculated using a rolling average with windows of size 5 kb, including only SNPs at the same frequency as the core SNP in the average. In addition, we show the allele frequencies of the GWAS and a top-scoring β SNP in each representative population. (b) Approximate haplotype lengths for each population.

Hughes and Nei (1988). In particular, we found a strong signal in the HLA at a locus influencing response to Hepatitis B infection, rs3077 (Jiang *et al.*, 2015; DeGiorgio *et al.*, 2014; Thursz *et al.*, 1997). Several additional top sites in our scan matched those from DeGiorgio *et al.* (2014). These include sites that tag phenotypic associations (Welter *et al.*, 2014), such as *MYRIP*, involved with sleep-related phenotypes (Gottlieb *et al.*, 2007), and *BICC1*, associated with corneal astigmatism (Lopes *et al.*, 2013). We focus on two of our top-scoring regions, located in the *CADM2* and *WFS1* genes. In addition to passing the 1000 Genomes strict filter and the RepeatMasker test, these haplotypes also passed Hardy-Weinberg filtering **(Section 3.2)**.

3.4 A signature of balancing selection at the *CADM2* locus

One of our top-scoring regions fell within an intron of the cell adhesion molecule 2 gene, *CADM2*. This locus contains a haplotype with β scores falling in the top 0.25 percentile in 17 of the 1000 Genomes populations, and scoring in the top 0.75 percentile across all 26 populations **(Fig. 3.2)**. This site was also a top scoring SNP in the CEU population based on the $T2$ statistic (DeGiorgio *et al.*, 2014). In our scan using the folded β statistic, this haplotype contained top-scoring variants in 20 populations, indicating the result was not due to ancestral allele miscalling. In the remaining six populations, the haplotype was at folded frequency 0.15 or lower, where the folded version of β has significantly reduced power.

To elucidate the potential mechanisms contributing to the signal in this region, we overlapped multiple genomic datasets to identify potential functional variants that were tightly linked with our haplotype signature. First, one variant that perfectly tags (EUR $r^2 = 1.0$) our signature, rs17518584, has been genome-wide significantly associated with cognitive functions, including information processing speed (Davies *et al.*, 2015; Ibrahim-Verbaas *et al.*, 2016). Second, multiple variants in this region co-localized (EUR r^2 between $0.9 - 1$ with rs17518584) with eQTLs of *CADM2* in numerous tissues (Lung, Adipose, Skeletal Muscle, Heart-Left Ventricle), though notably not in brain (The GTEx Consortium, 2015). That said, several SNPs with regulatory potential (RegulomeDB scores of 3a or higher) are also strongly tagged

by our high-scoring haplotype (EUR r^2 between $0.9 - 1.0$ with rs17518584), which include regions of open chromatin in Cerebellum and other cell types (Boyle *et al.*, 2012). Several SNPs on this haplotype, particularly rs1449378 and rs1449379, fall in enhancers in several brain tissues, including the hippocampus (Ernst and Kellis, 2012; Boyle *et al.*, 2012). Taken collectively, these data suggest that our haplotype tags a region of regulatory potential that may influence the expression of *CADM2*, and potentially implicates cognitive or neuronal phenotypes in the selective pressure at this site.

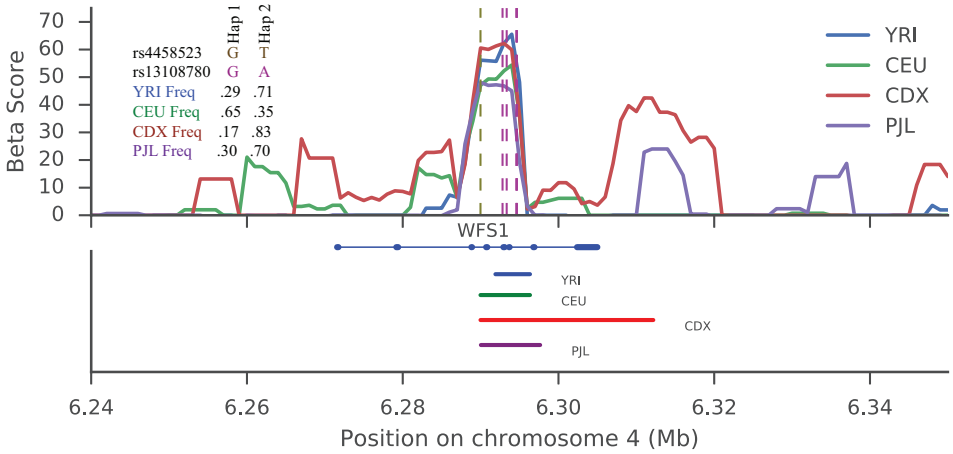

Figure 3.3: Signal of balancing selection at the WFS1 gene. (a) rs4458523 is the lead GWAS SNP for diabetes, and is marked by the brown vertical dashed line. The purple dashed line marks 5 regulatory variants found on the balanced haplotype. In addition, we show the allele frequencies of the GWAS and a top-scoring β SNP in each representative population. (b) Approximate haplotype lengths for each population.

3.5 A signature of balancing selection near the diabetes associated locus, *WFS1*

We identified a novel region of interest within the intron of *WFS1*, a transmembrane glycoprotein localized primarily to the endoplasmic reticulum (ER). *WFS1* functions in protein assembly (Takei *et al.*, 2006) and is an important regulator of the unfolded protein and ER Stress Response pathways (Fonseca *et al.*, 2005). A haplotype in this region (approximately 3.5 kb) contains approximately 26 variants, 3 of which are in high-quality windows and are high-scoring β in all populations **(Fig. 3.3)**. The haplotype was also in the top 1 percentile in our folded β scan in 21 populations. In the remaining 5 populations, this haplotype was at frequency 0.82 or higher, where the folded version of β has significantly lower power than the unfolded version.

Our identified high-scoring haplotype tags several functional and phenotypic variant associations. First, one variant that perfectly tags our signature (EUR $r^2 = 1.0$), rs4458523, has been previously associated with type 2 diabetes (Voight *et al.*, 2010; Mahajan *et al.*, 2014). Second, multiple variants in this region are associated with expression-level changes of *WFS1* in numerous tissues (The GTEx Consortium, 2015); these variants are strongly tagged by our high-scoring haplotype (EUR r^2 between $0.85 - 0.9$ with rs4458523). Finally, several SNPs with regulatory potential (RegulomeDB scores of 2b or higher) are also strongly tagged by our high-scoring haplotype (EUR r^2 between $0.9-1.0$ with rs4458523). Taken collectively, these data suggest that our haplotype tags a region of strong regulatory potential that is likely to influence

the expression of *WFS1*.

3.6 Discussion of top β loci

When overlapping our top β scores with lead GWAS SNPs from the GWAS catalog, we discover over 50 potentially balanced loci that have phenotypic associations **(Table 3.1)**. These include plausibly selected phenotypes, including asthma, schizophrenia and age at menarche. However, more work is needed to discover how these loci may be influencing these phenotypes. Furthermore, it is very difficult to know for sure whether balancing selection has been acting at the putatively balanced loci, as any statistic for balancing selection has a non-zero false positive rate. I also note that GWAS results from one population may not be applicable to another (Martin *et al.*, 2017), adding further complexity.

Although it is impossible to know the true selective pressure underlying our highlighted loci, our results suggest that balancing selection could contribute to the genetic architecture of complex traits in human populations. At the *CADM2* locus, functional genomics data suggests that our haplotype signature may connect to brain-related biology. Intriguingly, a recent report also noted a strong signature of selection at this locus in canine (Freedman *et al.*, 2016), suggesting a possibility of convergent evolution. That said, the phenotypes that have resulted in a historical fitness trade-off at this locus are far from obvious.

Similarly, speculation on the potential phenotypes subject to balancing selection at *WFS1* should also be interpreted cautiously. It is known that autosomal recessive, loss of function mutations in this gene cause Wolfram Syndrome. This gene is a component of the unfolded protein response Fonseca *et al.* (2005) and is involved with ER maintenance in pancreatic β-cells. Furthermore, deficiency of *WFS1* results in increased ER stress, impairment of cell cycle, and ultimately increased apoptosis of beta-cells Yamada *et al.* (2006). These data would suggest that reduced expression of *WFS1* would be diabetes risk increasing; however, eQTLs that co-localized with the diabetes risk-increasing allele *elevate* expression, at least in non-pancreas tissue, suggesting perhaps a more complex functional mechanism. Furthermore, how the unfolded protein response could connect to historical balancing selection is also not immediately obvious. One possibility derives from recent work suggesting that these pathways respond not only to stimulus from nutrients or ER stress, but also to pathogens Nakamura *et al.* (2010). This could suggest the possibility that expression of *WFS1* is optimized in part to respond to pathogen exposure at a population level.

GWAS SNP	Phenotype	Reported Gene(s)	Pubmed ID
rs17110736	Dialysis-related mortality	ABCA4	21546767
rs1478912	Response to taxane treatment	RYR2	23006423
rs78037194	Schizophrenia	NR	26198764
rs12120588	Urate levels in overweight individuals	SPATA17	25811787
rs3771166	Asthma	IL18R1	20860503
rs9807989	Asthma	IL18R1	22561531
rs11568377	Systolic blood pressure in sickle cell anemia	ABCB11	24058526
rs9287719	Prostate cancer	NOL10	25217961
rs7577463	Schizophrenia	NR	26198764
rs6599077	Sleep-related phenotypes	MYRIP	17903308
rs9861887	Visceral/subcutaneous adipose tissue ratio	CNTN6	22589738
rs17518584	Information processing speed	CADM2	25869804
rs17518584	Cognitive function	CADM2	25644384
rs1801214	Type 2 diabetes	WFS1	20581827
rs11942476	IgG glycosylation	NR	23382691
rs4458523	Type 2 diabetes	WFS1	24509480
rs1967256	Response to antipsychotic treatment	GPR98	20195266
rs3077	Hepatitis B	HLA-DPA1	21750111
rs365302	Coronary heart disease	FNDC1	21606135
rs10947262	Knee osteoarthritis	HLA, BTNL2	20305777
rs3077	Hepatitis B (viral clearance)	HLA-DPA1	22737229
rs10947261	Crohns disease	HLA, BTNL2	23850713
rs3077	Chronic hepatitis B infection	HLA-DP	23760081
rs12196860	Psychosis (atypical)	COL21A1	24132900
rs10447419	PR interval	intergenic	23534349
rs3077	Chronic hepatitis B infection	HLA-DPA1	25802187
rs1747593	Sitting height ratio	NR	25865494
rs2349775	Neuroticism	NXPH1	18762592
rs7804356	Type 1 diabetes	intergenic	19430480
rs10486158	Bipolar disorder and schizophrenia	NR	20889312
rs10486483	Crohns disease	intergenic	23128233
rs10486483	Crohns disease	NR	26192919
rs10486483	Inflammatory bowel disease	NR	26192919
rs2294008	Bladder cancer	PSCA	19648920
rs2294008	Bladder cancer	PSCA	20972438
rs2294008	Duodenal ulcer	PSCA	22387998
rs7818688	Vincristine-induced peripheral neuropathy	NDUFAF6	25710658
rs2294008	Gastric cancer	PSCA	26098866
rs2294008	Gastric adenocarcinoma	PSCA	26098866
rs7044529	Central corneal thickness	COL5A1	20719862
rs1927702	Body mass index	NR	19851299
rs7044529	Corneal structure	COL5A1	23291589

rs7084402	Refractive error	BICC1	23396134
rs1658442	Corneal astigmatism	NR	23322567
rs17134585	Blood metabolite ratios	AKR1C4	24816252
rs1832007	Triglycerides	AKR1C4	24097068
rs59132240	Ejection fraction in T. cruzi seropositivity	NR	24324551
rs10846260	Bone mineral density and age at menarche	DERA	26312577
rs1926657	Breast cancer	ABCC4	17903305
rs6563739	Menarche (age at onset)	COG6	25231870
rs6574644	Obesity-related traits	STON2	23251661
rs17111396	Uric acid levels	TSHR	21294900
rs607541	Obesity-related traits	SQRDL	23251661
rs11071033	Menarche (age at onset)	UNC13C	23599027
rs7165042	Myocardial infarction	ADAMTS7	26343387
rs4468572	Coronary artery disease	ADAMTS7	26343387
rs8070723	Parkinsons disease	MAPT	21044948
rs8070723	Progressive supranuclear palsy	MAPT	21685912
rs12185268	Parkinsons disease	MAPT	21738487
rs9303525	Intracranial volume	MAPT, GRN, CRHR1, STH	22504418
rs12373124	Male-pattern baldness	intergenic	22693459
rs892961	Airflow obstruction	SEPT9	22837378
rs1864325	Bone mineral density	MAPT	22504420
rs17577094	Parkinsons disease	MAPT	24842889
rs17649553	Parkinsons disease	MAPT	25064009
rs1981997	Interstitial lung disease	MAPT	23583980
rs12185268	Corticobasal degeneration	MAPT	26077951
rs8072451	Subcortical brain region volumes	MAPT, GRN, CRHR1, STH	25607358
rs17689882	Subcortical brain region volumes	CRHR1	25607358
rs11876941	Body mass index (interaction)	DCC	23192594
rs2281135	Liver enzyme levels	PNPLA3, SAMM50	18940312
rs2896019	Hematological, biochemical traits	PARVB, PNPLA3, SAMM50	20139978
rs2896019	Non-alcoholic fatty liver disease	PARVB, PNPLA3, SAMM50	23535911

Table 3.1: Lead SNPs from the GWAS catalog that are in high linkage disequilibrium ($r^2 > 0.9$) with a top β SNP.

Chapter 4

Detecting ancient balancing selection using substitutions

The results of this chapter are presented in:

Siewert, K. M. and Voight, B. F. 2018. BetaScan2: Standardized statistics to detect balancing selection utilizing substitution data. *bioRxiv*: 497255.

The $\beta^{(1)}$ statistics use only polymorphism (i.e. within-species mutation) data to detect selection. However, balancing selection also reduces the number of substitutions (see section 1.2.2). This chapters details an addition to the β suite of statistics, $\beta^{(2)}$, which looks not only at polymorphism data, but also substitution data. This statistic has increased power over $\beta^{(1)}$ to detect balancing selection.

4.1 Derivation of $\hat{\theta}_D$ and its variance

We first derive our estimator of the mutation rate based on the divergence between two species, $\hat{\theta}_D$. To measure divergence, we use the number of substitutions, which we define as the nucleotide positions in which the outgroup individual is different than all ingroup individuals. We note that this differs from the measure of between-species divergence in the HKA test, which is instead the average number of differences between a randomly selected ingroup and outgroup gamete (Hudson *et al.*, 1987). We choose this measure of divergence, because as noted in Hudson *et al.* (1987), it has slightly lower variance than the one they used (albeit, a slightly more complex derivation). We assume that there is a single outgroup individual and that the time since speciation is sufficiently long that the ingroup coalescence occurred prior to coalescing with the outgroup. Throughout our derivations, we assume Hardy-Weinberg equilibrium, an infinite sites model, and no recombination. In practice, recombination will act to decrease the variance, making our standardized β statistics conservative (Tajima, 1989).

We model divergence using the coalescence tree of the ingroup individuals and the outgroup individuals **(Fig. 4.1)**. This tree contains two parts that can contribute to substitutions. Considering the tree backward in time, these parts are (i) after the coalescence of the common ancestor of the ingroup and outgroup individuals and (ii) after the coalescence within each species, but before coalescence between the two species.

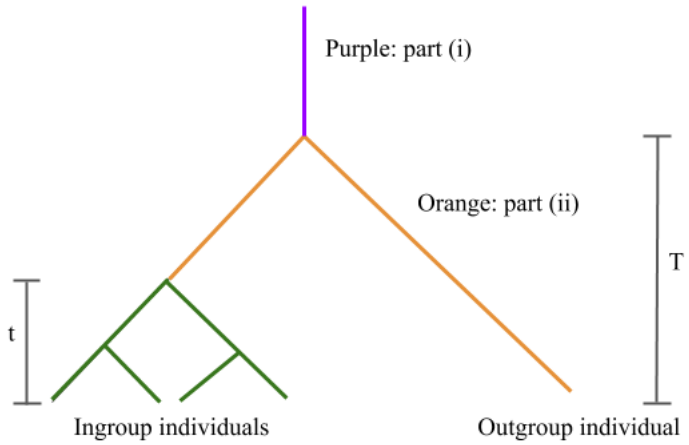

Figure 4.1: The coalescent tree between two species can be broken up into three segments. Here, mutations on the green segment result in polymorphisms, the orange part results in substitutions and the purple part results in shared derived alleles between the two species. t is the ingroup speciation time and T is the coalescence time between the two species.

Expected number and variance of substitutions from part (ii), $D_{(ii)}$:

The number of substitutions in part (ii) of the tree is Poisson distributed, $D_{(ii)} \sim Poisson(\mu L)$, where μ is the mutation rate and L is the branch length in part (ii). The branch length is given by $L \approx 2T - t$, where T is the coalescence time of the ingroup and outgroup (i.e. speciation time), and t is the coalescence time of the ingroup species. The expected value and variance of t is given in Tavaré (1984). Let F be the Poisson distributed variable representing the number of mutations along these branches, and G represent the height of the ingroup coalescence tree. Define the surveyed size of sampled chromosomes to be n, the effective population size as N_e, and $\theta = 4N_e\mu$ with μ as the usual mutation rate per base per generation. We can derive the expected value and variance of $D_{(ii)}$ using the properties of compound probability distributions, the theorem for moments of the height of a coalescent tree

from Tavaré (1984) and the mean and variance of the Poisson distribution. First, we find the expected value:

$$\begin{aligned} E[D_{(ii)}] &= E_G[E_F[D_{(ii)}|L]] \\ &= E_G[\mu(2T-t)] \\ &= 2T\mu - E_G[t]\mu \\ &= 2T\mu - 4N_e\big(1-\frac{1}{n}\big)\mu \\ &= \theta\Big(\frac{T}{2N_e} - (1-\frac{1}{n})\Big) \end{aligned} \tag{4.1.1}$$

Next, we find the variance of $D_{(ii)}$, using the variance of t from Tavaré (1984):

$$\begin{aligned} Var[D_{(ii)}] &= E_G[Var_F[D_{(ii)}|L]] + Var_G[E_F[D_{(ii)}|L]] \\ &= E_G[(2T-t)\mu] + Var_G[(2T-t)\mu] \\ &= 2T\mu - E_G[t]\mu + Var_G[2T\mu - t\mu] \\ &= 2T\mu - 4N_e(1-\frac{1}{n})\mu + \mu^2 Var_G[t] \\ &= 2T\mu - 4N_e\mu + \frac{4N_e\mu}{n} + (4N_e\mu)^2\Big(\sum_{i=2}^{n}\frac{1}{i^2(i-1)^2}\Big) \end{aligned} \tag{4.1.2}$$

Expected number and variance of substitutions from part (i), $D_{(i)}$:

We obtain the expected number and variance of $D_{(i)}$ by noting that it is equivalent

to the difference between two random gametes from Watterson (1975).

$$E[D_{(i)}] = \theta \tag{4.1.3}$$

$$Var[D_{(i)}] = \theta + \theta^2 \tag{4.1.4}$$

Expected number and variance of substitutions from whole tree:

We denote the total number of substitutions as $D = D_{(i)} + D_{(ii)}$. From Eqs. (1) and (3) above, the expected value is then given by:

$$\begin{aligned} E[D] &= E[D_{(i)}] + E[D_{(ii)}] \\ &= \theta + \theta\Big(\frac{T}{2N_e} - (1 - \frac{1}{n})\Big) \\ &= \theta\Big(\frac{T}{2N_e} + \frac{1}{n}\Big) \end{aligned} \tag{4.1.5}$$

Because the coalescent process in part (i) is independent of the coalescence process in part (ii), we can simply add variances from Eqs. (2) and (4) above to obtain:

$$\begin{aligned} Var[D] &= Var[D_i] + Var[D_{ii}] \\ &= \theta + \theta^2 + 2T\mu - 4N_e\mu + \frac{4N_e\mu}{n} + (4N_e\mu)^2(\sum_{i=2}^{n} \frac{1}{i^2(i-1)^2}) \\ &= \theta^2 + \frac{T\theta}{2N_e} + \frac{\theta}{n} + \theta^2 \sum_{i=2}^{n} \frac{1}{i^2(i-1)^2} \end{aligned} \tag{4.1.6}$$

We note that our results for the mean and variance of D are simplified forms of equations 29C and 31C from Hey (1991) when taking the large T limit.

Solving for θ in Eq. 4.1.5 we obtain $\hat{\theta}_D$:

$$\hat{\theta}_D = \frac{D}{\frac{T}{2N_e} + \frac{1}{n}} \tag{4.1.7}$$

The variance of $\hat{\theta}_D$ is then:

$$\begin{aligned} Var[\hat{\theta}_D] &= Var\left[\frac{D}{\frac{T}{2N_e} + \frac{1}{n}}\right] \\ &= \left(\frac{1}{\frac{T}{2N_e} + \frac{1}{n}}\right)^2 Var[D] \\ &= \left(\frac{1}{\frac{T}{2N_e} + \frac{1}{n}}\right)^2 \left(\theta^2 + \frac{T\theta}{2N_e} + \frac{\theta}{n} + \theta^2 \sum_{i=2}^{n} \frac{1}{i^2(i-1)^2}\right) \end{aligned} \tag{4.1.8}$$

This leads to:

$$\beta^{(2)}_{std} = \frac{\beta^{(2)}}{\sqrt{Var[\beta^{(2)}]}} = \frac{\hat{\theta}_\beta - \hat{\theta}_D}{\sqrt{Var[\hat{\theta}_\beta] + Var[\hat{\theta}_D]}} \tag{4.1.9}$$

where T is the estimated speciation time in generations, N_e is the estimated effective population size of the ingroup species, and $\hat{\theta}$ is the estimated mutation rate. For simplicity, we assume that Cov[$\hat{\theta}_\beta$, $\hat{\theta}_D$]=0. This assumption results in a slight underestimate of the variance of $\beta^{(2)}$, as would be expected due to the small negative

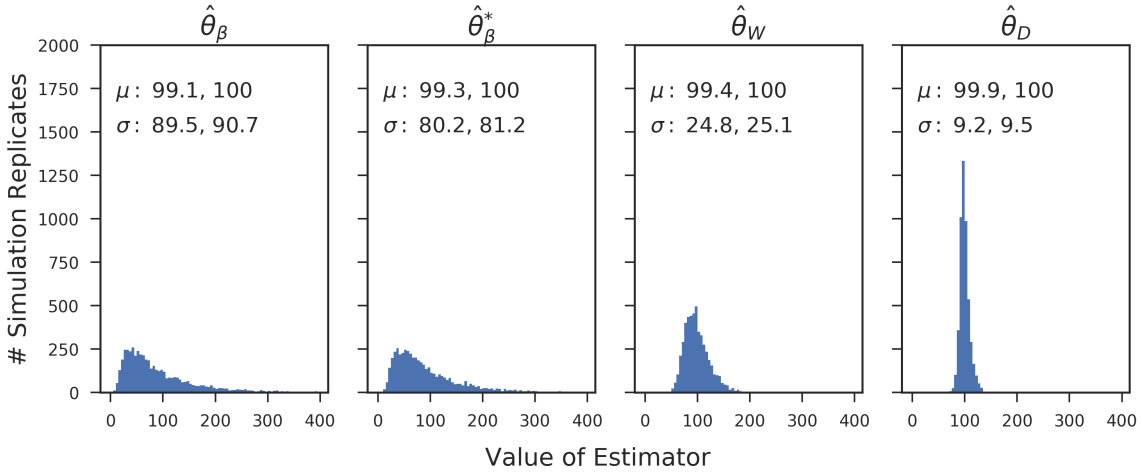

Figure 4.2: Distribution of each θ estimator on simulated 100kb windows with no selection or recombination and equilibrium demography. Core frequencies were chosen to be 0.5, regardless of whether a SNP of that frequency was found in the window. Mean (μ) and standard deviation (σ) are displayed, with the first number being the sample value, and the second being the theoretical value.

covariance between the ingroup coalescence time and the number of substitutions. However, under an equilibrium model (constant population size), the expected values and variances of θ_β,θ_W and θ_D fit those seen in simulations, confirming these derivations as sounds approximations. Furthermore, the mean of each β statistic is approximately zero, as would be expected, and the variance is extremely close to what would be expected. **(Fig. 4.2, 4.3)**.

4.2 Estimation of the speciation time

The variance of $\beta^{(2)}$ is also dependent on the speciation time (in coalescent units, i.e. units of $2N_e$). The speciation time can be obtained from prior demographic analyses of the species of interests, or by estimating it from the data at hand. The software

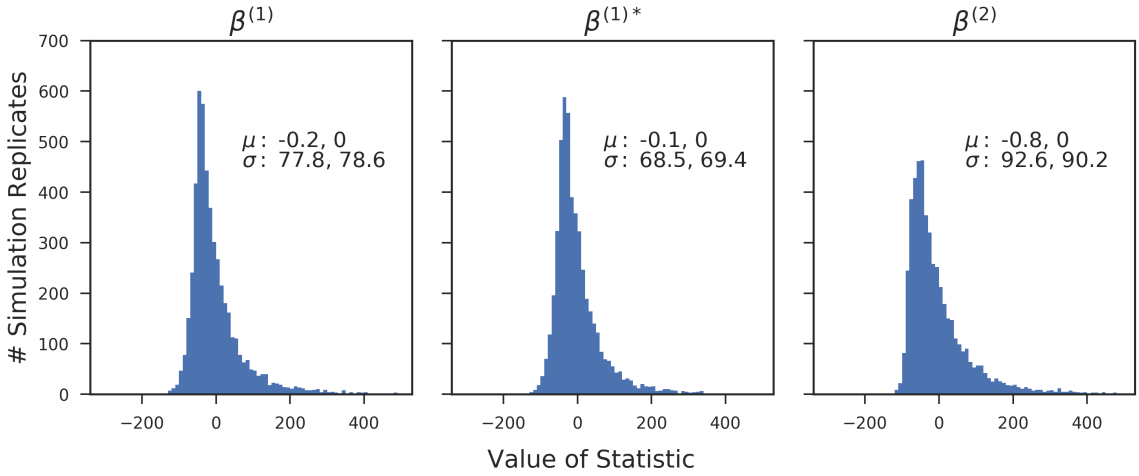

Figure 4.3: Distribution of each β statistic on simulated 100kb windows with no selection or recombination and equilibrium demography. Core frequencies were chosen to be 0.5, regardless of whether a SNP of that frequency was found in the window. Mean (μ) and standard deviation (σ) are displayed, with the first number being the sample value, and the second being the theoretical value.

presented in DeGiorgio *et al.* (2014) implements an estimator of divergence based on the site frequency spectrum and the number of substitutions, which we recommend when prior estimates of speciation time are not available. However, the power of β is very robust to choice of T **(Fig. 4.4)**.

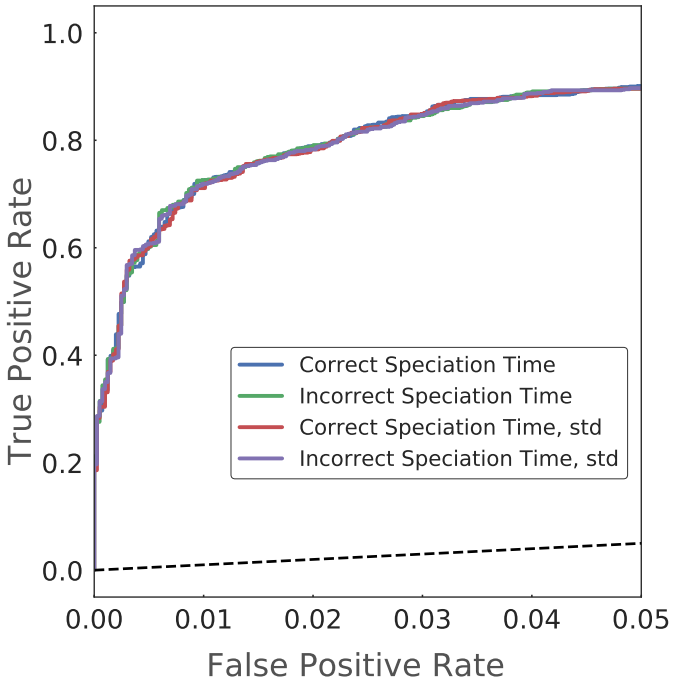

Figure 4.4: Power of $\beta^{(2)}$ and $\beta^{(2)}_{std}$ when the speciation time parameter is correctly specified as 250,000 generations prior to sampling versus when it is underestimated by 100,000 generations. An equilibrium frequency of 50% and a selection age of 250,000 generations prior to sampling were used.

4.3 Power analysis

4.3.1 Power analysis of $\beta^{(2)}$ and standardized β statistics

The power analyses in chapter 2 focus on the $\beta^{(1)}$ statistics. In order to evaluate the power of $\beta^{(2)}$, we repeated these analyses. We find that $\beta^{(2)}$ has higher power than either $\beta^{(1)}$ statistic, demonstrating that substitution counts provide additional signal over polymorphism data **(Figs. 4.5a, 4.6, 4.7)**. When there is mutation rate variation across simulations, we find that standardization improves power **(Fig. 4.5b)**.

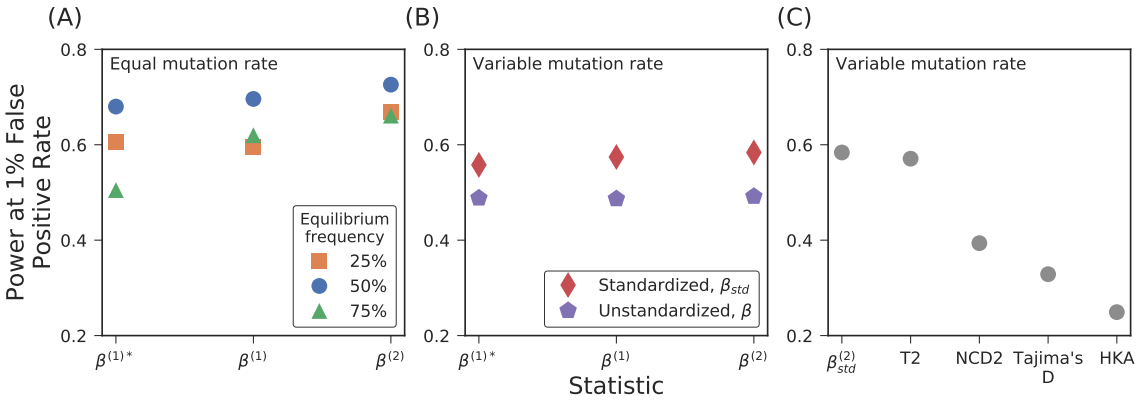

Figure 4.5: Power of β statistics at (A) different equilibrium frequencies and (B) with mutation rate variation, where one half of neutral and balanced simulation replicates had a mutation rate of 2.5×10^{-8} (our default rate), and the remaining half had a rate of 1.25×10^{-8}. (C) Power of $\beta^{(2)}$ compared to other methods for detecting balancing selection. An equilibrium frequency of 50% was used for (B) and (C). The values of each statistic were compared between simulations containing only neutral variants (True Negatives) or with a balanced variant (True Positives).

Next, we compare power to alternative methods: $NCD2$ (Bitarello *et al.*, 2018), NCD_{mid} (Cheng and DeGiorgio, 2018), $T1$ and $T2$ (DeGiorgio *et al.*, 2014), Tajima's D (Tajima, 1989) and the HKA test (Hudson *et al.*, 1987). When there is mutation rate variation, we find that $\beta^{(2)}_{std}$ performs the strongest **(Fig. 4.5c)**. However, $T2$, a method that relies on grids of simulations to generate composite likelihoods, performs best when the mutation rate is stable and the parameters underlying the simulations for $T2$ match the selection scenario, as is the case with the older balancing selection category. When there is a mismatch between these, the power of $T2$ is reduced, and β starts to outperform **(Figs. 4.6, 4.7)**. Our results were not biased by window size **(Fig. 4.8)**. As discussed below, the relative performance of these methods stayed consistent using both power analysis paradigms that have been used in the literature **(Figs. 4.6, 4.7)**.

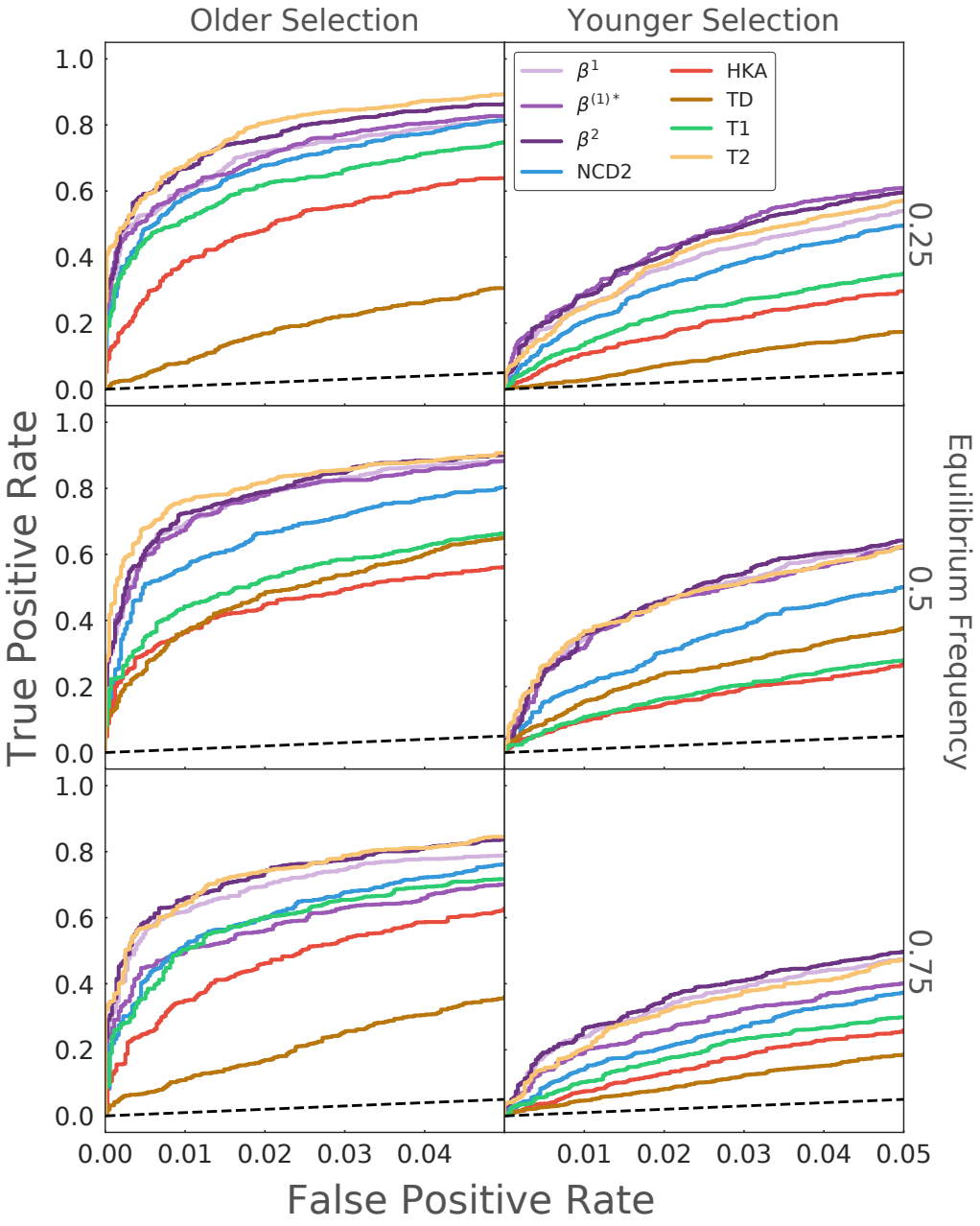

Figure 4.6: Power of methods to detect long-term balancing selection. Power was calculated based on simulation replicates containing only neutral variants (True Negatives) or containing a balanced variant that was introduced (True Positives). The score of the balanced SNP was used for each statistic, as was the score of a SNP from the neutral simulations matched for frequency. Rows correspond to simulations of balanced alleles at equilibrium frequencies 0.25, 0.50, and 0.75. Columns correspond to older and more recent selection, beginning 250,000 and 100,000 generations prior to sampling, respectively. The black line goes from the origin to a true and false positive rate of one, and would correspond to a method with no discriminatory power.

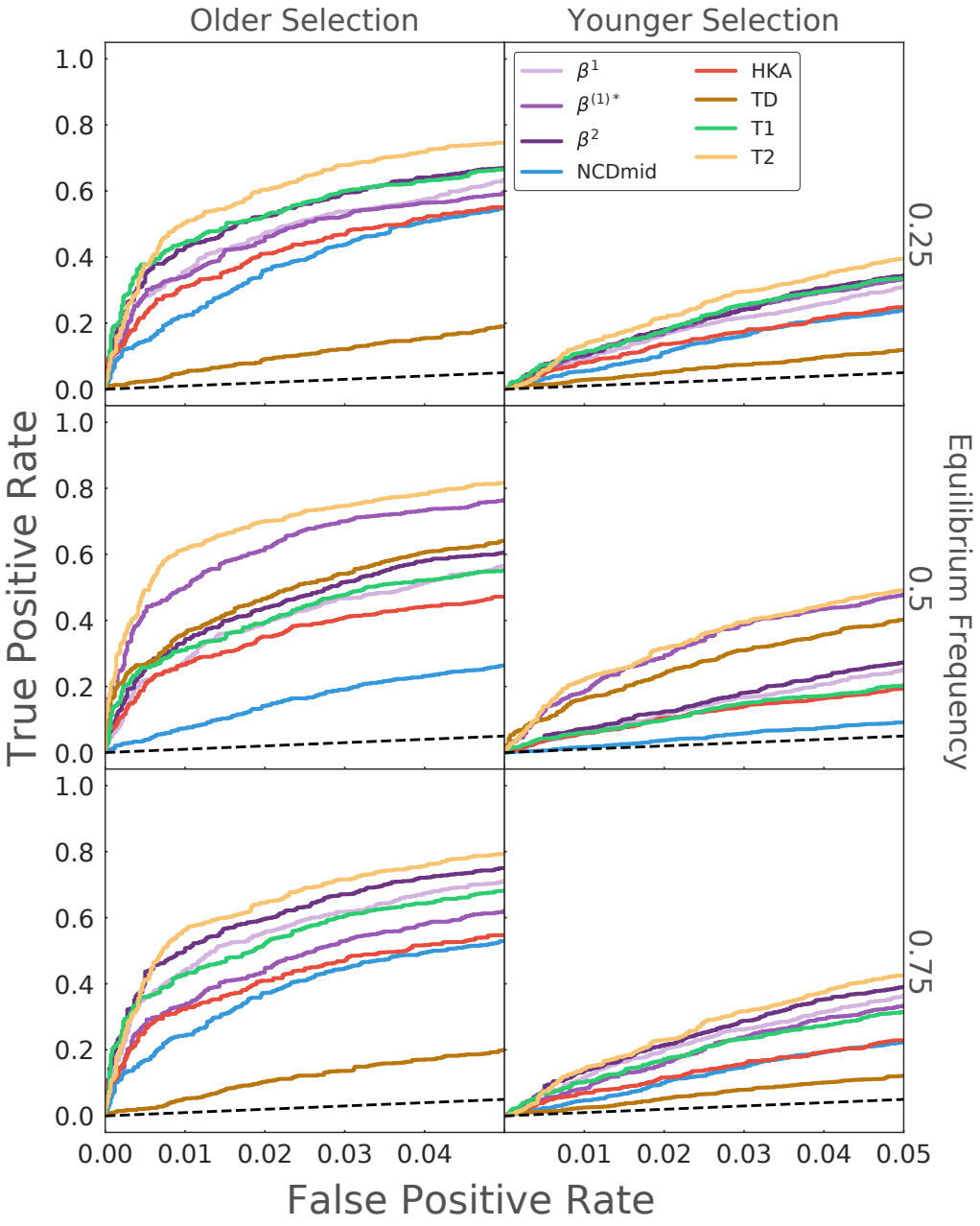

Figure 4.7: Power of methods to detect ancient balancing selection without matching for allele frequency. Power was calculated based on simulation replicates containing only neutral variants (True Negatives) or containing a balanced variant that was introduced (True Positives). The maximum value of each statistic in each simulated 10kb window was used. Rows correspond to simulations of balanced alleles at equilibrium frequencies 0.25, 0.50, and 0.75. Columns correspond to older and more recent selection, beginning 250,000 and 100,000 generations prior to sampling, respectively.

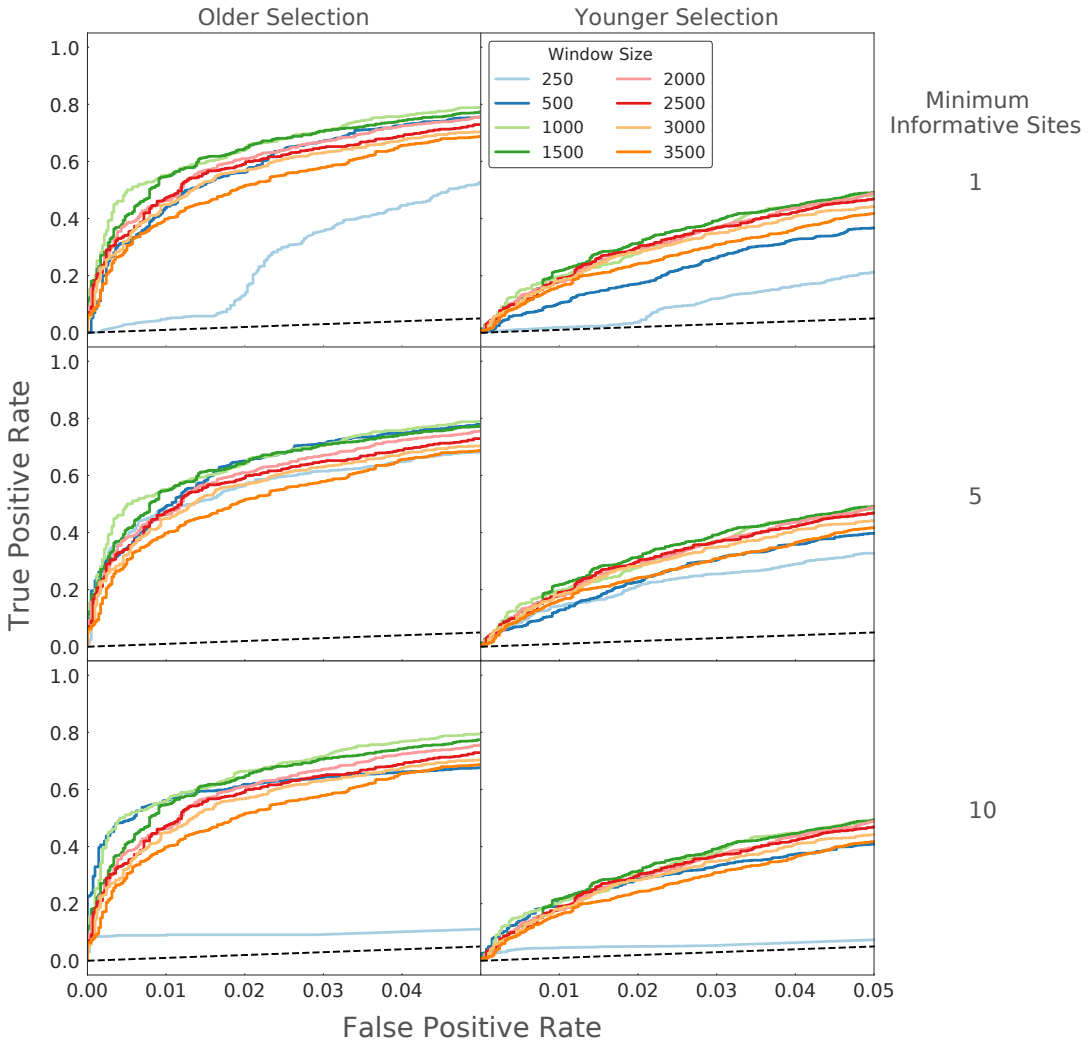

Figure 4.8: Power of the $NCD2$ statistic using different window sizes and minimum number of informative sites (SNPs plus substitutions). Here, we show that the 1kb window we used for $NCD2$ is optimal for power. In addition, Bitarello *et al.* (2018) suggested that $NCD2$ may require a minimum of 5 or 10 informative sites for maximum power. We show that when using an optimal window size, no minimum is needed under our simulation parameters. For each row, any windows with less than the given number of informative sites were called as neutral. Units are in base pairs. An equilibrium frequency of 50% was used. Columns correspond to older and more recent selection, beginning 250,000 and 100,000 generations prior to sampling, respectively.

4.3.2 Techniques for power comparison

Two techniques for power analysis have been used in the literature. We use both, and show that the relative performance of the various methods remains roughly consistent

across comparison methods.

I *Single target frequency.* This power comparison method answers the question "How well do the methods distinguish between a balanced SNP at a certain frequency and a neutral SNP of a similar frequency?"

To perform this power analysis, we directly scored the simulated balanced SNP for each statistic. We note that the balanced SNPs are usually not at exactly the equilibrium frequency, due to genetic drift and sub-sampling of individuals, but most are within 10% of the equilibrium frequency (data not shown). For this reason, for the neutral scores, we found a SNP in each neutral simulation replicate within frequency 10% of the equilibrium frequency of the balanced SNPs, and used its score. If there was not a SNP within frequency 10%, we did not use that simulation replicate in that power analysis. In this way, all methods are testing for allelic class build-up at approximately one frequency. Instead of using the frequency of the core SNP, the $NCD2$ statistic requires the user to specify a target frequency (Cheng and DeGiorgio, 2018)(Bitarello *et al.*, 2018). We used a value equal to the expected equilibrium frequency in the balanced simulations, which represents the best case for NCD. This power comparison method was used for all figures except **(Fig. 4.7)**.

II *Multiple target frequencies.* This power comparison method answers the question "How well do the methods distinguish between a window with a balanced SNP and any window without a balanced SNP?"

The maximum value of each statistic is used in each simulated balanced and neutral window. However, the values of $NCD2$ calculated using different target frequencies are not comparable. To address this issue, Cheng and DeGiorgio (2018) developed NCD_{mid}, which calculates a modified NCD2 value using a grid of target frequencies. We compare the power of NCD using NCD_{mid} for this comparison type. This power comparison method was used for Fig. 4.7).

4.3.3 Comparison with prior power analyses.

In both **Fig. 4.6** and **Fig. 4.7**, our results coincide with those of Cheng and DeGiorgio (2018) in that $T2$ outperforms $NCD2$ or NCD_{mid}. However Cheng and DeGiorgio (2018) did not compare the power of β, as they were focused on methods to detect selection shared between multiple species, for which β is not especially tailored.

The performance of the β, $T1$, and $T2$ statistics relative to $NCD2$ deviate from that found in Bitarello *et al.* (2018). This is due to two differences between our power analysis methods. The first, as pointed out in Cheng and DeGiorgio (2018), is that 100, not 10, informative sites were used to calculate $T1$ and $T2$ in Bitarello *et al.* (2018), reducing the power of $T1$ and $T2$. Like in Cheng and DeGiorgio (2018), we used 10 informative sites so the window sizes for all statistics are as equivalent as possible.

Secondly, in Bitarello *et al.* (2018), the value of NCD2 used to calculate power is based

on a single core/target SNP frequency. In contrast, the T and β values that were used were the maximum T or β score across all SNPs in the window. Because $T2$ and β adapts to use the frequency of each SNP it is calculated on, this is equivalent to using power comparison method (I) for NCD2, but power comparison method (II) for the other statistics. This increases the number of core SNP frequencies that the β and T statistics must test, and therefore artificially increase the false positive rate relative to NCD2. When we use the same power comparison method for all statistics, whether it be using a single target/core frequency or all allele frequencies in the simulated window, we find that $T2$ and the β statistics tend to perform the strongest **(Fig. 4.6, Fig. 4.7)**. $T2$ uses simulated site frequency spectra under balancing selection and neutrality. When the computational power, outgroup sequence and knowledge of demographic parameters exist to perform these simulations, our findings suggest that $T2$ may be the ideal statistic to use, while the β statistics may be best to use otherwise.

4.3.4 Comparison of $\beta^{(2)}$ and $NCD2$ statistics.

The $NCD2$ statistic measures the average frequency difference between SNPs in a window and a target frequency, with substitutions considered as SNPs of frequency 0.

$$NCD(tf) = \sqrt{\frac{\sum_{i=1}^{n}(p_i - tf)^2}{n}} \tag{4.3.1}$$

where p_i is the allele frequency of the ith of n SNPs in a window. The target frequency is analogous to the core SNP frequency from β or $T1/T2$. However, NCD requires this parameter to be set by the user, unlike with β or $T1/T2$, which use the frequency of the SNP at the center of each window as the target/core SNP frequency. The reason for this is that the expected value of NCD is not constant across target frequencies, so NCD scores can only be compared to scores using the same target frequency.

$NCD2$ and $\beta^{(2)}$ are similar in their approach, in that they both explicitly capture excessive allele frequency correlation. We posit that the relative strength of $\beta^{(2)}$ compared to $NCD2$ is due to several factors. The first is that by using a difference of two unbiased estimators of the mutation rate, $\beta^{(2)}$ has a constant expected value (zero), whereas the expected value of $NCD2$ varies with target frequency (Bitarello *et al.*, 2018). This enables β values to be compared across different allele frequencies, so that it can use the exact frequency of the core SNP, instead of having to use the same target frequency across all SNPs.

Secondly, instead of taking the square of the *average* frequency difference between each SNP and the target frequency, β is a function of the *sum* of the frequency similarity. This means that SNPs at large frequency differences away from the core site frequency have very little effect on β. In contrast, for NCD, these SNPs factor into the average and add noise. For instance, NCD with a target frequency of 50 will return the same value if there is a window with ten SNPs at frequency 50% and ten singletons as it will in a window with two SNPs at frequency 50% and two singletons. In contrast, the β score will be nearly five times higher in the first case

than the second as there is a five times stronger signal of allelic class build-up. In addition, rare variant calls are often problematic in real data and are not indicative of the presence or absence of balancing selection, so an ideal statistic would not be influenced by their presence.

Thirdly, because it does not take into account speciation time like $T2$ or $\beta^{(2)}$ does, the distribution of NCD2 is heavily dependent on speciation time. Too long of speciation time would increase the number of substitutions considerably and could dwarf a signal in the polymorphism portion of the spectrum.

Lastly, NCD2 considers substitutions to be SNPs of frequency zero, which can cause a false signal of excessive allele frequency correlation when trying to detect balancing selection at extreme equilibrium frequencies (Cheng and DeGiorgio, 2018). We instead consider them in a separate estimator, $\hat{\theta}_D$.

4.4 Estimation of the background mutation rate

Calculating the variance of each β statistic requires knowledge of the underlying mutation rate. We recommend estimating this from the data. Several estimators of the mutation rate may be appropriate. If sequencing errors are expected to be rare, then Watterson's estimator is a good choice, as it has very low variance. However, in practice rare variants can be prone to false or missing calls. In situations like this, estimators which ignore rare variation may be a better choice. Achaz (2008)

proposed an estimator similar to Watterson's estimator which uses the number of segregating sites, singletons excluded, to estimate the mutation rate: $\hat{\theta}_{S_{-\xi_1}} = \frac{S_{-\xi_1}}{a_n - 1}$, where $S_{-\xi_1}$ is the number of segregating sites excluding singletons and $a_n = \sum_{i=1}^{n} \frac{1}{n-1}$. Achaz (2008) introduces a similar estimator which excludes singletons for when only a folded site frequency spectrum is available. One of these two estimators are more appropriate in case where singletons are prone to false positive or missing calls (e.g., elevated error rates from technology or low-pass sequencing coverage).

The mutation rate can either be estimated at a genome-wide level or for individual loci. Estimating at a locus-by-locus level allows the variance to reflect changes in mutability or background selection. However, doing so can also increase the variance of the mutation rate estimator, as the size of window used to estimate the mutation rate will be smaller. If too small of a window is used, the variance of the denominator of the standardized statistics may swamp signals of selection from the numerator, decreasing power. In practice, we recommend using the largest window you think still reflects local changes in mutation rate that will be important. Using simulations of human parameters, we find that Watterson's theta on 1kb windows surrounding the core SNP does a poor job of estimating the background mutation rate (data not shown), while 10kb windows do significantly better **(Fig. 4.9)**.

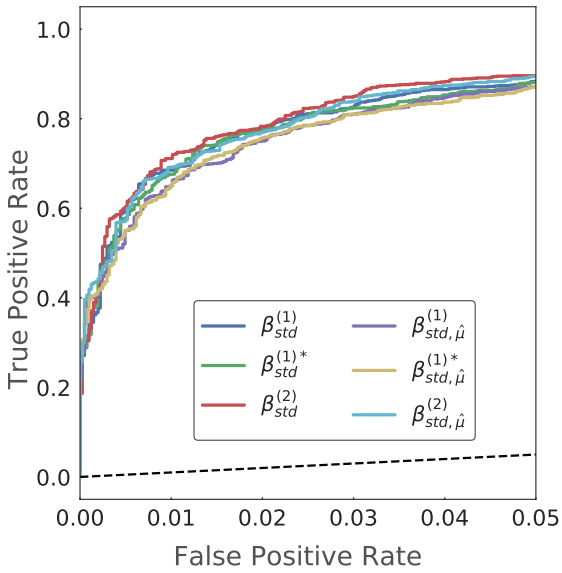

Figure 4.9: Power of β statistics when the background mutation rate is estimated using a 10kb window centered at the core SNP ($\hat{\mu}$) versus using the true mutation rate. A mutation rate of 2.5×10^{-8}, an equilibrium frequency of 50% and a selection age of 250,000 generations prior to sampling was used.

Chapter 5

Conclusions and future directions

5.1 The β statistic in perspective

Here, I have described our new suite of methods for detecting balancing selection: the β statistics. It is my hope that due to their power and flexibility, others will find the β statistics useful. To this aim, I have implemented these statistics into a toolkit, and provided an extensive user manual, to allow others to scan the genome of their species of interest: https://github.com/ksiewert/BetaScan. This toolkit is also quick – calculating β on an entire chromosome takes less than a minute.

The β statistics are the first of a new class of summary statistics to detect balancing selection using a more precise signature of balancing selection than Tajima's D: an excess number of SNPs at near-identical frequencies to the balanced alleles. It is

illustrative to compare the power of $T2$ and $\beta^{(2)}$. The $T2$ statistic of DeGiorgio *et al.* (2014) uses a simulated site frequency spectrum to generate a composite likelihood of seeing a SNP of a given frequency at each distance from a balanced SNP, conditioned on the balanced SNP being at the observed equilibrium frequency. Prior to the β statistics, $T2$ had significantly higher power than any other method to detect balancing selection. The similarity in power between $T2$ and the β statistics indicate the signature explicitly captured by the β statistics may encompass most of the the signature which $T2$ implicitly captures using simulated site frequency spectra.

The closest classic method to β in both statistical structure and signature captured is Tajima's D (Tajima, 1989). Both statistics look for an excess of SNPs around a given frequency, and use a difference between an estimator sensitive to this excess and Watterson's θ to detect selection. In the case of Tajima's D, an excess of SNPs near frequency 50% is measured, as this is when heterozygosity, and therefore θ_π, is highest. If the equilibrium frequency of the balanced allele is close to 50%, then Tajima's D has power closer to $T2$ or the β statistics. However, if the equilibrium frequency is at a more extreme frequency, the power of Tajima's D suffers, as SNPs fixed in alleleic class will not result in as high of heterozygosity as if they were at frequency 50%. Unlike Tajima's D, the β statistics utilize the observed SNP frequencies to adjust which frequencies should be tested for an excess of in order to detect balancing selection, increasing power.

Furthermore, Tajima's D does not use an explicit similarity function, although it in effect weights SNPs an amount proportional to the square of their similarity to 50%.

In contrast, β weights SNPs an amount proportion to the power p (which would be the square if $p = 2$) of their fraction of maximum possible allele frequency similarity to the core SNP. The higher power of $\beta^{(1)*}$ over Tajima's D even at equilibrium frequency 50% indicates that this similarity function used by β more precisely captures the shape of the peak in the site frequency spectrum caused by balancing selection than heterozygosity does.

5.2 Potential improvements to the β statistics

Many possible improvements to this class of statistics are possible. For instance, β currently uses a fixed window size around each core SNP. However, it may be possible to infer the optimal window size from looking at patterns of linkage disequilibrium at the locus. The optimal window would contain the balanced haplotype, but would not extend past it. However, this potential future direction has the downside that it would likely have increased run-time and memory usage, as it would require looking at individual-level sequence data and making inferences about the length of haplotypes from that data. In this way, it would lose some of the advantages of summary statistics and approach coalescent estimators in its complexity.

An additional increase in power could result from a method which does not use the same allele frequency correlation function across all SNPs in a window. For instance, SNPs closest to the core SNP are more likely to have not experienced recombination

between them and the core SNP, and therefore are more likely to be at exactly the core SNP frequency. Therefore, a larger value of the p parameter may be more suitable. A better theoretical understanding of the signature of balancing selection could inform a more optimally designed measure of allele frequency similarity. Along the same lines, additional theoretical development on the signature of balancing selection with recombination could inform how allele frequency similarity is expected to change across distances from the balanced SNP.

A promising class of methods for detecting selection are coalescent estimators (see section 1.3.6). Currently, it is difficult to apply these methods to data or to compare their power to existing methods, because of high computational cost. However, as methods to estimate coalescent times improve, they may become a standard approach to detect selection.

5.3 Large-scale effects of balancing selection on the genome: future avenues for exploration

In this thesis, I describe only the application of $\beta^{(1)}$ to detect selection in humans. However, the power of $\beta^{(1)}$ is decreased in regions of low mutation rate or higher background selection. Therefore, it is of lower power in regions with the highest probability of containing functional mutations which may be balanced. β_{std} allows one to scan the genome in a manner less affected by mutation rate. A scan using a standardized β statistic would allow a better characterization of the effects of balanc-

ing selection genome-wide. I would expect that such a scan would find a much higher overlap between top β SNPs and trans-species SNPs, and also a stronger enrichment for top β scores near genes.

However, a key challenge of quantifying the effects of balancing selection genome-wide remains that the null distribution of any statistic we use to detect selection is unknown under human demography. Therefore, the appropriate significance threshold and corresponding false discovery rate is unknown, making it difficult to measure how common ancient balancing selection has been in the evolutionary history of a species of interest. Simulations may be performed to generate an empirical p-value threshold, however if these simulations do not accurately model demography, they may do a poor job generating a simulated null distribution.

Perhaps an approach similar to the one taken to measure the genome-wide effects of positive selection could be taken. To overcome this challenge when investigating the frequency of positive selection, evidence for correlations between signatures of positive selection, such as population differentiation or haplotype length, and functional annotations, such as distance to nearest gene or PolyPhen annotation have been quantified (Hernandez *et al.*, 2011; Enard *et al.*, 2014). If positive selection is indeed prevalent throughout evolution, then you would expect correlation between these features. A similar approach could be taken for balancing selection. Key to this analysis would be a well-powered statistic for detecting the signature of balancing selection. In addition, these statistics must not be confounded by factors such as mutation rate, which could induce spurious correlation between features. For this

reason, statistics with known variance, such as β, may be a good choice.

From revealing selective pressures and resulting adaptations, to increasing our understanding of the mutation load, future research on balancing selection promises to increase our understanding of evolution and genetic architecture. The β_{std} statistics presented in this thesis enable researchers to better answer these questions through high powered detection of balancing selection.

Bibliography

Achaz, G. 2008. Testing for neutrality in samples with sequencing errors. *Genetics*, 179(3): 1409–24.

Achaz, G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1): 249–58.

Aidoo, M., Terlouw, D. J., Kolczak, M. S., *et al.* 2002. Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet*, 359(9314): 1311–1312.

Allison, A. C. 1954. Protection afforded by sickle-cell trait against subtertian malareal infection. *British medical journal*, 1(4857): 290–4.

Anderson, T. M., VonHoldt, B. M., Candille, S. I., *et al.* 2009. Molecular and evolutionary history of melanism in North American gray wolves. *Science*, 323(5919): 1339–43.

Andres, A. M., Hubisz, M. J., Indap, A., *et al.* 2009. Targets of Balancing Selection in the Human Genome. *Molecular Biology and Evolution*, 26(12): 2755–2764.

Asthana, S., Schmidt, S., and Sunyaev, S. 2005. A limited role for balancing selection. *Trends in Genetics*, 21(1): 30–32.

Bersaglieri, T., Sabeti, P. C., Patterson, N., *et al.* 2004. Genetic signatures of strong recent positive selection at the lactase gene. *American journal of human genetics*, 74(6): 1111–20.

Bitarello, B. D., de Filippo, C., Teixeira, J. C., *et al.* 2018. Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biology and Evolution*, 10(3): 939–955.

Boyle, A. P., Hong, E. L., Hariharan, M., *et al.* 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, 22(9): 1790–1797.

Bruce, A. B. 1910. The Mendelian theory of heredity and the augmentation of vigor.

Bubb, K. L., Bovee, D., Buckley, D., *et al.* 2006. Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics*, 173(4): 2165–77.

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4): 379–384.

Cheng, X. and DeGiorgio, M. 2018. Detection of shared balancing selection in the absence of trans-species polymorphism. *Molecular Biology and Evolution*.

Crow, J. F. 1987. Muller, Dobzhansky, and overdominance. *Journal of the History of Biology*, 20(3): 351–380.

Crow, J. F. 1998. 90 Years Ago: The Beginning of Hybrid Maize. *Genetics*, 148(3): 923–928.

Danecek, P., Auton, A., Abecasis, G., *et al.* 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–2158.

Darwin, C. 1878. *The Effects of Cross and Self Fertilization in the Vegetable Kingdom*.

Davies, G., Armstrong, N., Bis, J. C., *et al.* 2015. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53,949). *Molecular psychiatry*, 20(2): 183–192.

De Boer, R. J., Borghans, J. A. M., van Boven, M., Kemir, C., and Weissing, F. J. 2004. Heterozygote advantage fails to explain the high degree of polymorphism of the MHC. *Immunogenetics*, 55(11): 725–731.

DeGiorgio, M., Lohmueller, K. E., and Nielsen, R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS genetics*, 10(8): e1004561.

Do, R., Balick, D., Li, H., *et al.* 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics*, 47(2): 126–131.

East, E. M. 1936. Heterosis. *Genetics*, 21(4).

Enard, D., Messer, P. W., and Petrov, D. A. 2014. Genome-wide signals of positive selection in human evolution. *Genome research*, 24(6): 885–95.

Ernst, J. and Kellis, M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3): 215–216.

Ewens, W. J. and Thomson, G. 1970. Heterozygote selective advantage. *Annals of Human Genetics*, 33(4): 365–376.

Eyre-Walker, A. and Keightley, P. D. 1999. High genomic deleterious mutation rates in hominids. *Nature*, 397(6717): 344–347.

Fay, J. C. and Wu, C. I. 2000. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3): 1405–13.

Ferretti, L., Klassmann, A., Raineri, E., *et al.* 2018. The neutral frequency spectrum of linked sites. *Theoretical Population Biology*, 123: 70–79.

Fonseca, S. G., Fukuma, M., Lipson, K. L., *et al.* 2005. WFS1 is a novel component of the unfolded protein response and maintains homeostasis of the endoplasmic reticulum in pancreatic beta-cells. *The Journal of biological chemistry*, 280(47): 39609–39615.

Freedman, A. H., Schweizer, R. M., Ortega-Del Vecchyo, D., *et al.* 2016. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. *PLOS Genetics*, 12(3): e1005851.

Fu, Y. X. 1995. Statistical Properties of Segregating Sites. *Theoretical Population Biology*, 48(2): 172–197.

Gao, Z., Przeworski, M., and Sella, G. 2014. Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution*, 69(2): 431–46.

Gottlieb, D. J., O'Connor, G. T., and Wilk, J. B. 2007. Genome-wide association of sleep and circadian phenotypes. *BMC Medical Genetics*, 8(Suppl 1): S9.

Haldane, J.B.S., Jayakar, S. 1963. Polymorphism due to selection of varying direction. *Journal of Genetics*, 58: 237–242.

Haller, B. C. and Messer, P. W. 2017. SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Molecular biology and evolution*, 34(1): 230–240.

Hedrick, P. W. 1998. Balancing selection and MHC. *Genetica*, 104(3): 207–214.

Hedrick, P. W. 2012. What is the evidence for heterozygote advantage selection? *Trends in Ecology & Evolution*, 27(12): 698–704.

Hedrick, P. W., Ginevan, M. E., and Ewing, E. P. 1976. Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics*, 7: 1–32.

Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G., and Gravel, S. 2015. Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16(6): 333–343.

Hernandez, R. D., Kelley, J. L., Elyashiv, E., *et al.* 2011. Classic selective sweeps were rare in recent human evolution. *Science*, 331(6019): 920–4.

Hey, J. 1991. The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. *Genetics*, 128(4).

Hudson, R. R. and Kaplan, N. L. 1988. The coalescent process in models with selection and recombination. *Genetics*, 120(3): 831–40.

Hudson, R. R., Kreitman, M., and Aguadé, M. 1987. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics*, 116(1): 153–159.

Hughes, A. L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186): 167–170.

Hull, F. H. 1945. Recurrent selection and specific combining ability in corn. *Journal of the American Society of Agronomy*, 37: 134–145.

Hull, F. H. 1946. Overdominance and corn breeding where hybrid seed is not feasible when. *Journal of the American Society of Agronomy*, 38: 1100–1103.

Ibrahim-Verbaas, C. A., Bressler, J., Debette, S., *et al.* 2016. GWAS for executive function and processing speed suggests involvement of the CADM2 gene. *Molecular Psychiatry*, 21(2): 189–197.

Ilardo, M. A., Moltke, I., Korneliussen, T. S., *et al.* 2018. Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell*, 173(3): 569–580.e15.

Ingvarsson, P. K. 2004. Population subdivision and the Hudson-Kreitman-Aguade test: testing for deviations from the neutral model in organelle genomes. *Genetical research*, 83(1): 31–9.

Jiang, D.-K., Ma, X.-P., Yu, H., *et al.* 2015. Genetic variants in five novel loci including CFB and CD40 predispose to chronic hepatitis B. *Hepatology*, 62(1): 118–128.

Kamberov, Y. G., Wang, S., Tan, J., *et al.* 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell*, 152(4): 691–702.

Kaplan, N., Darden, T., and Hudson, R. 1988. The coalescent process in models with selection. *Genetics*, 120(3): 819–29.

Keinan, A. and Clark, A. G. 2012. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science*, 336(6082): 740–743.

Key, F. M., Teixeira, J. C., de Filippo, C., and Andrés, A. M. 2014. Advantageous diversity maintained by balancing selection in humans. *Current opinion in genetics & development*, 29C: 45–51.

Klassmann, A. and Ferretti, L. 2018. The third moments of the site frequency spectrum. *Theoretical Population Biology*, 120: 16–28.

Leffler, E. M., Gao, Z., Pfeifer, S., *et al.* 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, 339(6127): 1578–82.

Lenz, T. L., Spirin, V., Jordan, D. M., and Sunyaev, S. R. 2016. Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. *Molecular Biology and Evolution*, 33(10): 2555–2564.

Levene, H. 1953. Genetic Equilibrium When More Than One Ecological Niche is Available. *The American Naturalist*, 87(836): 331–333.

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357): 493–496.

Lopes, M. C., Hysi, P. G., Verhoeven, V. J. M., *et al.* 2013. Identification of a Candidate Gene for Astigmatism. *Investigative Opthalmology & Visual Science*, 54(2): 1260.

Luzzatto, L. 2012. Sickle cell anaemia and malaria. *Mediterranean journal of hematology and infectious diseases*, 4(1): e2012065.

Mahajan, A., Go, M. J., Zhang, W., *et al.* 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3): 234–244.

Martin, A. R., Gignoux, C. R., Walters, R. K., *et al.* 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics*, 100(4): 635–649.

Moll, R. H., Lindsey, M. F., and Robinson, H. F. 1963. Estimates of genetic variances and level of dominance in maize. *Genetics*, 49(3): 411–423.

Nakamura, T., Furuhashi, M., Li, P., *et al.* 2010. Double-stranded RNA-dependent protein kinase links pathogen sensing with stress and metabolic homeostasis. *Cell*, 140(3): 338–48.

Network, M. G. E. 2015. A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*, 526(7572): 253–257.

Osier, M., Pakstis, A. J., Kidd, J. R., *et al.* 1999. Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. *American journal of human genetics*, 64(4): 1147–57.

Palamara, P. F., Terhorst, J., Song, Y. S., and Price, A. L. 2018. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50(9): 1311–1317.

Piel, F. B., Patil, A. P., Howes, R. E., *et al.* 2010. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature communications*, 1: 104.

Rasmussen, M. D., Hubisz, M. M. J. M. J., Gronau, I., *et al.* 2014. Genome-wide inference of ancestral recombination graphs. *PLoS genetics*, 10(5): e1004342.

Robertson, A. 1962. Selection for heterozygotes in small populations. *Genetics*, 47(9): 1291–1300.

Sano, E. B., Wall, C. A., Hutchins, P. R., and Miller, S. R. 2018. Ancient balancing selection on heterocyst function in a cosmopolitan cyanobacterium. *Nature Ecology & Evolution*, 2(3): 510–519.

Schierup, M. H., Vekemans, X., and Charlesworth, D. 2000. The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetical research*, 76(1): 51–62.

Schweizer, R. M., Durvasula, A., Smith, J., *et al.* 2018. Natural Selection and Origin of a Melanistic Allele in North American Gray Wolves. *Molecular Biology and Evolution*, 35(5): 1190–1209.

Ségurel, L., Thompson, E. E., Flutre, T., *et al.* 2012. The ABO blood group is a trans-species polymorphism in primates.

Shull, G. H. 1948. What is heterosis? *Genetics*, 33(5): 439–446.

Siewert, K. M. and Voight, B. F. 2017. Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Molecular Biology and Evolution*, 34(11): 2996–3005.

Siewert, K. M. and Voight, B. F. 2018. BetaScan2: Standardized statistics to detect balancing selection utilizing substitution data. *bioRxiv*.

Simons, Y. B., Turchin, M. C., Pritchard, J. K., and Sella, G. 2014. The deleterious mutation load is insensitive to recent population history. *Nature Genetics*, 46(3): 220–224.

Slade, R. W. and McCallum, H. I. 1992. Overdominant vs. frequency-dependent selection at MHC loci. *Genetics*, 132(3): 861–4.

Speidel, L., Forest, M., Shi, S., and Myers, S. 2019. A method for genome-wide genealogy estimation for thousands of samples. *bioRxiv*, page 550558.

Tajima, F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3): 585.

Takahata, N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proceedings of the National Academy of Sciences*, 87(7).

Takahata, N. and Nei, M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*.

Takei, D., Ishihara, H., Yamaguchi, S., *et al.* 2006. WFS1 protein modulates the free Ca2+ concentration in the endoplasmic reticulum. *FEBS Letters*, 580(24): 5635–5640.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical population biology*, 26(2): 119–64.

Teixeira, J. C., de Filippo, C., Weihmann, A., *et al.* 2015. Long-Term Balancing Selection in LAD1 Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos. *Molecular biology and evolution*, 32(5): msv007–.

The 1000 Genomes Consortium 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68–74.

The GTEx Consortium 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235): 648–660.

Thursz, M. R., Thomas, H. C., Greenwood, B. M., and Hill, A. V. 1997. Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nature Genetics*, 17(1): 11–12.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., *et al.* 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics*, 39(1): 31–40.

Vernot, B. and Akey, J. M. 2015. Complex history of admixture between modern humans and Neandertals. *American journal of human genetics*, 96(3): 448–453.

Voight, B. F., Scott, L. J., Steinthorsdottir, V., *et al.* 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics*, 42(7): 579–589.

Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2): 256–276.

Welter, D., MacArthur, J., Morales, J., *et al.* 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(Database issue): 1001–1006.

Wheat, C. W., Haag, C. R., Marden, J. H., Hanski, I., and Frilander, M. J. 2010. Nucleotide Polymorphism at a Gene (Pgi) under Balancing Selection in a Butterfly Metapopulation. *Molecular Biology and Evolution*, 27(2): 267–281.

Whitfield, J. B. 2002. Alcohol dehydrogenase and alcohol dependence: variation in genotype-associated risk between populations. *American journal of human genetics*, 71(5): 1247–50; author reply 1250–1.

Yamada, T., Ishihara, H., Tamura, A., *et al.* 2006. WFS1-deficiency increases endoplasmic reticulum stress, impairs cell cycle progression and triggers the apoptotic pathway specifically in pancreatic -cells. *Human Molecular Genetics*, 15(10): 1600–1609.